THE INTERNET ENCYCLOPEDIA

Volume 3

Edited by Hossein Bidgoli

THE INTERNET ENCYCLOPEDIA

Volume 3
P–Z

Hossein Bidgoli
Editor-in-Chief
California State University
Bakersfield, California

John Wiley & Sons, Inc.

THE INTERNET ENCYCLOPEDIA

Volume 3
P–Z

Hossein Bidgoli
Editor-in-Chief
California State University
Bakersfield, California

John Wiley & Sons, Inc.

This book is printed on acid-free paper. ♾

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.
Published simultaneously in Canada.

Library of Congress Cataloging-in-Publication Data:

The Internet encyclopedia / edited by Hossein Bidgoli.
p. cm.
Includes bibliographical references and index.
ISBN 0-471-22202-X (CLOTH VOL 1 : alk. paper) – ISBN 0-471-22204-6 (CLOTH VOL 2 : alk. paper) – ISBN 0-471-22203-8 (CLOTH VOL 3 : alk. paper) – ISBN 0-471-22201-1 (CLOTH SET : alk. paper)
1. Internet–Encyclopedias. I. Bidgoli, Hossein.
TK5105.875.I57I5466 2003
004.67′8′03–dc21

2002155552

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

To so many fine memories of my brother, Mohsen, for his uncompromising belief in the power of education.

About the Editor-in-Chief

Hossein Bidgoli, Ph.D., is Professor of Management Information Systems at California State University. Dr. Bidgoli helped set up the first PC lab in the United States. He is the author of 43 textbooks, 27 manuals, and over four dozen technical articles and papers on various aspects of computer applications, e-commerce, and information systems, which have been published and presented throughout the world. Dr. Bidgoli also serves as the editor-in-chief of *Encyclopedia of Information Systems*.

Dr. Bidgoli was selected as the California State University, Bakersfield's 2001–2002 Professor of the Year.

Editorial Board

Contents

Volume 2

Volume 3

Chapter List by Subject Area

Wireless Internet and E-commerce

Contributors

Tarek Abdelzaher
University of Virginia
Web Quality of Service

Charles Abzug
James Madison University
Linux Operating System

Patricia Adams
Education Resources
Strategic Alliances

Carol A. Akerelrea
Colorado State University
Usability Testing: An Evaluation Process for Internet Communications

Gary C. Anders
Arizona State University West
Online Auctions

Amy W. Apon
University of Arkansas
Public Networks

Pierre A. Balthazard
Arizona State University West
Groupware

Ashok Deo Bardhan
University of California, Berkeley
Real Estate

Joey Bargsten
University of Oregon
Multimedia

Hossein Bidgoli
California State University, Bakersfield
Computer Literacy
Internet Literacy

Gerald Bluhm
Tyco Fire & Security
Patent Law

Robert J. Boncella
Washburn University
Secure Sockets Layer (SSL)

J. Efrim Boritz
University of Waterloo, Canada
XBRL (Extensible Business Reporting Language): Business Reporting with XML

Sviatoslav Braynov
State University of New York at Buffalo
Data Mining in E-commerce
Personalization and Customization Technologies

Randy M. Brooks
Millikin University
Online Publishing

Colleen Brown
Purdue University
History of the Internet

Tara Brown-L'Bahy
Harvard University
Distance Learning (Virtual Learning)

Linda S. Bruenjes
Lasell College
Internet2

Gerard J. Burke
University of Florida
Supply Chain Management

L. Jean Camp
Harvard University
Peer-to-Peer Systems

Charles J. Campbell
The University of Memphis
Java

Janice E. Carrillo
University of Florida
Inventory Management

Michael A. Carrillo
Oracle Corporation
Inventory Management

Lillian N. Cassel
Villanova University
Wireless Application Protocol (WAP)

J. Cecil
New Mexico State University
Virtual Enterprises

Haluk Cetin
Murray State University
Geographic Information Systems (GIS) and the Internet

Henry Chan
The Hong Kong Polytechnic University, China
Consumer-Oriented Electronic Commerce

C. Janie Chang
San José State University
Public Accounting Firms

Camille Chin
West Virginia University
Cybercrime and Cyberfraud

T. Matthew Ciolek
The Australian National University, Australia
Online Religion

Timothy W. Cole
University of Illinois at Urbana-Champaign
Visual Basic Scripting Edition (VBScript)

Fred Condo
California State University, Chico
Cascading Style Sheets (CSS)

David E. Cook
University of Derby, United Kingdom
Standards and Protocols in Data Communications

Marco Cremonini
Università di Milano, Italy
Disaster Recovery Planning

Mary J. Cronin
Boston College
Mobile Commerce
Jaime J. Dávila
Hampshire College
Digital Divide
Chris Dede
Harvard University
Distance Learning (Virtual Learning)
Victoria S. Dennis
Minnesota State Bar Association
Law Firms
Lynn A. DeNoia
Rensselaer Polytechnic Institute
Wide Area and Metropolitan Area Networks
Nikhilesh Dholakia
University of Rhode Island
Gender and Internet Usage
Global Diffusion of the Internet
Ruby Roy Dholakia
University of Rhode Island
Gender and Internet Usage
Global Diffusion of the Internet
Vesna Dolnicar
University of Ljubljana, Slovenia
Benchmarking Internet
Rich Dorfman
WebFeats! and Waukesha County Technical College
Extensible Markup Language (XML)
Magda El Zarki
University of California—Irvine
Wireless Internet
Larry P. English
Information Impact International, Inc.
Information Quality in Internet and E-business Environments
Roman Erenshteyn
Goldey-Beacom College
ActiveX
Ray Everett-Church
ePrivacy Group, Inc.
Privacy Law
Trademark Law
Patrick J. Fahy
Athabasca University, Canada
Web-Based Training
Gerald R. Ferrera
Bentley College
Copyright Law
Daniel R. Fesenmaier
University of Illinois at Urbana–Champaign
Travel and Tourism
C. Patrick Fleenor
Seattle University
Feasibility of Global E-business Projects
Marcia H. Flicker
Fordham University
Securities Trading on the Internet
Immanuel Freedman
Dr. Immanuel Freedman, Inc.
Video Compression
Borko Furht
Florida Atlantic University
Interactive Multimedia on the Web
Jayne Gackenbach
Athabasca University, Canada
Health Issues
Alan Gaitenby
University of Massachusetts, Amherst
Online Dispute Resolution
Bruce Garrison
University of Miami
Online News Services (Online Journalism)
G. David Garson
North Carolina State University
E-government
Roger Gate
IBM United Kingdom Ltd., United Kingdom
Electronic Funds Transfer
Mario Giannini
Code Fighter, Inc., and Columbia University
C/C++
Julia Alpert Gladstone
Bryant College
International Cyberlaw
Mary C. Gilly
University of California, Irvine
Consumer Behavior
Robert H. Goffman
Concordia University
Electronic Procurement
James E. Goldman
Purdue University
Firewalls
Sven Graupner
Hewlett-Packard Laboratories
Web Services
Robert H. Greenfield
Computer Consulting
Circuit, Message, and Packet Switching
Ulrike Gretzel
University of Illinois at Urbana–Champaign
Travel and Tourism
Paul Gronke
Reed College
Politics
Jim Grubbs
University of Illinois at Springfield
E-mail and Instant Messaging
Mohsen Guizani
Western Michigan University
Wireless Communications Applications
Jon Gunderson
University of Illinois at Urbana–Champaign
Universally Accessible Web Resources: Designing for People with Disabilities
Babita Gupta
California State University, Monterey Bay
Global Issues
Louisa Ha
Bowling Green State University
Webcasting

Kirk Hallahan
Colorado State University
Online Public Relations
Diane M. Hamilton
Rowan University
Business-to-Consumer (B2C) Internet Business Models
Robert W. Heath Jr.
The University of Texas at Austin
Digital Communication
Geert Heijenk
University of Twente, The Netherlands
Wireless Internet
Jesse M. Heines
University of Massachusetts Lowell
Extensible Stylesheet Language (XSL)
Rodney J. Heisterberg
Notre Dame de Namur University and
Rod Heisterberg Associates
Collaborative Commerce (C-Commerce)
Steven J. Henry
Wolf, Greenfield & Sacks, P.C.
Open Source Development and Licensing
Julie Hersberger
University of North Carolina at Greensboro
Internet Censorship
Kenneth Einar Himma
University of Washington
Legal, Social, and Ethical Issues
Matthias Holweg
Massachusetts Institute of Technology
Managing the Flow of Materials Across the Supply Chain
Russ Housley
Vigil Security, LLC
Public Key Infrastructure (PKI)
Yeong-Hyeon Hwang
University of Illinois at Urbana–Champaign
Travel and Tourism
Robert E. Irie
SPAWAR Systems Center San Diego
Web Site Design
Linda C. Isenhour
University of Central Florida
Human Resources Management
Hans-Arno Jacobsen
University of Toronto, Canada
Application Service Providers (ASPs)
Charles W. Jaeger
Southerrn Oregon University
Cyberterrorism
Dwight Jaffee
University of California, Berkeley
Real Estate
Sushil Jajodia
George Mason University
Intrusion Detection Techniques
Mark Jeffery
Northwestern University
Return on Investment Analysis for E-business Projects
Andrew Johnson
University of Illinois at Chicago
Virtual Reality on the Internet: Collaborative Virtual Reality
Ari Juels
RSA Laboratories
Encryption
Bhushan Kapoor
California State University, Fullerton
ActiveX Data Objects (ADO)
Joseph M. Kayany
Western Michigan University
Internet Etiquette (Netiquette)
Doug Kaye
RDS Strategies LLC
Web Hosting
Chuck Kelley
Excellence In Data, Inc.
Data Warehousing and Data Marts
Diane Ketelhut
Harvard University
Distance Learning (Virtual Learning)
Chang-Su Kim
Seoul National University, Korea
Data Compression
Wooyoung Kim
University of Illinois at Urbana-Champaign
Web Services
Jerry Kindall
Epok Inc.
Digital Identity
Brad Kleindl
Missouri Southern State University–Joplin
Value Chain Analysis
Graham Knight
University College London, United Kingdom
Internet Architecture
Craig D. Knuckles
Lake Forest College
DHTML (Dynamic HyperText Markup Language)
Jim Krause
Indiana University
Enhanced TV
Peter Kroon
Agere Systems
Speech and Audio Compression
Gary J. Krug
Eastern Washington University
Convergence of Data, Sound, and Video
Nir Kshetri
University of North Carolina
Gender and Internet Usage
Global Diffusion of the Internet
C.-C. Jay Kuo
University of Southern California
Data Compression
Stan Kurkovsky
Columbus State University
Common Gateway Interface (CGI) Scripts
Pamela M. H. Kwok
Hong Kong Polytechnic University, China
Wireless Marketing
Jennifer Lagier
Hartnell College
File Types

Thomas D. Lairson
Rollins College
Supply Chain Management and the Internet
Gary LaPoint
Syracuse University
International Supply Chain Management
Haniph A. Latchman
University of Florida
Managing a Network Environment
John LeBaron
University of Massachusetts Lowell
Internet2
Kenneth S. Lee
University of Pennsylvania
Wireless Internet
Jason Leigh
University of Illinois at Chicago
Virtual Reality on the Internet: Collaborative Virtual Reality
Margarita Maria Lenk
Colorado State University
Guidelines for a Comprehensive Security System
Nanette S. Levinson
American University
Developing Nations
Edwin E. Lewis Jr.
Johns Hopkins University
E-business ROI Simulations
David J. Loundy
DePaul University
Online Stalking
Robert H. Lowson
University of East Anglia, United Kingdom
E-systems for the Support of Manufacturing Operations
Supply Networks: Developing and Maintaining Relationships and Strategies
David Lukoff
Saybrook Graduate School and Research Center
Health Issues
Kuber Maharjan
Purdue University
Downloading from the Internet
Julie R. Mariga
Purdue University
Mobile Devices and Protocols
Mobile Operating Systems and Applications
Oge Marques
Florida Atlantic University
Interactive Multimedia on the Web
Prabhaker Mateti
Wright State University
TCP/IP Suite
Bruce R. Maxim
University of Michigan–Dearborn
Game Design: Games for the World Wide Web
Blayne E. Mayfield
Oklahoma State University
Visual C++ (Microsoft)
Cavan McCarthy
Louisiana State University
Digital Libraries
Patrick McDaniel
AT&T Labs
Authentication
David E. McDysan
WorldCom
Virtual Private Networks: Internet Protocol (IP) Based
Daniel J. McFarland
Rowan University
Client/Server Computing
Matthew K. McGowan
Bradley University
Electronic Data Interchange (EDI)
Nenad Medvidovic
University of Southern California
JavaBeans and Software Architecture
Nikunj R. Mehta
University of Southern California
JavaBeans and Software Architecture
John A. Mendonca
Purdue University
Organizational Impact
Weiyi Meng
State University of New York at Binghamton
Web Search Technology
Mark S. Merkow
E-commerce Guide
Secure Electronic Transactions (SET)
Mark Michael
King's College
HTML/XHTML (HyperText Markup Language/ Extensible HyperText Markup Language)
Physical Security
Brent A. Miller
IBM Corporation
Bluetooth™—A Wireless Personal Area Network
Robert K. Moniot
Fordham University
Software Piracy
Joseph Morabito
Stevens Institute of Technology
Online Analytical Processing (OLAP)
Roy Morris
Capitol College
Voice over Internet Protocol (IP)
Alec Nacamuli
IBM United Kingdom Ltd., United Kingdom
Electronic Funds Transfer
Annette Nellen
San José State University
Public Accounting Firms
Taxation Issues
Dale Nesbary
Oakland University
Nonprofit Organizations
Dat-Dao Nguyen
California State University, Northridge
Business-to-Business (B2B) Internet Business Models
Peng Ning
North Carolina State University
Intrusion Detection Techniques

Mark E. Nissen
Naval Postgraduate School
Intelligent Agents
Won Gyun No
University of Waterloo, Canada
XBRL (Extensible Business Reporting Language): Business Reporting with XML
Eric H. Nyberg
Carnegie Mellon University
Prototyping
Jeff Offutt
George Mason University
Software Design and Implementation in the Web Environment
Donal O'Mahony
University of Dublin, Ireland
Electronic Payment
Robert Oshana
Southern Methodist University
Capacity Planning for Web Services
Dennis O. Owen
Purdue University
Visual Basic
Raymond R. Panko
University of Hawaii at Manoa
Computer Security Incident Response Teams (CSIRTs)
Digital Signatures and Electronic Signatures
Internet Security Standards
Anand Paul
University of Florida
Inventory Management
Thomas L. Pigg
Jackson State Community College
Conducted Communications Media
Paul S. Piper
Western Washington University
Research on the Internet
Benjamin R. Pobanz
Purdue University
Mobile Devices and Protocols
Richard E. Potter
University of Illinois at Chicago
Groupware
Dennis M. Powers
Southern Oregon University
Cyberlaw: The Major Areas, Development, and Provisions
Paul R. Prabhaker
Illinois Institute of Technology
E-marketplaces
Etienne E. Pracht
University of South Florida
Health Insurance and Managed Care
Frederick Pratter
Eastern Oregon University
JavaServer Pages (JSP)
Robert W. Proctor
Purdue University
Human Factors and Ergonomics
Jian Qin
Syracuse University
Web Content Management
Zinovy Radovilsky
California State University, Hayward
Enterprise Resource Planning (ERP)
Jeremy Rasmussen
Sypris Electronics, LLC
Passwords
Peter Raven
Seattle University
Feasibility of Global E-business Projects
Amy W. Ray
Bentley College
Business Plans for E-commerce Projects
Julian J. Ray
Western New England College
Business-to-Business (B2B) Electronic Commerce
Pratap Reddy
Raritan Valley Community College
Internet Navigation (Basics, Services, and Portals)
Drummond Reed
OneName Corporation
Digital Identity
Vladimir V. Riabov
Rivier College
Storage Area Networks (SANs)
Nick Rich
Cardiff Business School, United Kingdom
Managing the Flow of Materials Across the Supply Chain
Malu Roldan
San Jose State University
Marketing Plans for an E-commerce Project
Constantine Roussos
Lynchburg College
JavaScript
Akhil Sahai
Hewlett-Packard Laboratories
Web Services
Eduardo Salas
University of Central Florida
Human Resources Management
Atul A. Salvekar
Intel Corp.
Digital Communication
Pierangela Samarati
Università di Milano, Italy
Disaster Recovery Planning
J. Christopher Sandvig
Western Washington University
Active Server Pages
Robert J. Schalkoff
Clemson University
Rule-Based and Expert Systems
Shannon Schelin
North Carolina State University
E-government
William T. Schiano
Bentley College
Intranets
Roy C. Schmidt
Bradley University
Risk Management in Internet-Based Software Projects

E. Eugene Schultz
University of California–Berkley Lab
Denial of Service Attacks
Windows 2000 Security
Steven D. Schwaitzberg
Tufts-New England Medical Center
Medical Care Delivery
Kathy Schwalbe
Augsburg College
Project Management Techniques
Mark Shacklette
The University of Chicago
Unix Operating System
P. M. Shankar
Drexel University
Propagation Characteristics of Wireless Channels
John Sherry
Purdue University
History of the Internet
Carolyn J. Siccama
University of Massachusetts Lowell
Internet2
Judith C. Simon
The University of Memphis
Java
Law Enforcement
Law Firms
Robert Simon
George Mason University
Middleware
Nirvikar Singh
University of California, Santa Cruz
Digital Economy
Clara L. Sitter
University of Denver
Library Management
Robert Slade
Consultant
Computer Viruses and Worms
Erick D. Slazinski
Purdue University
Structured Query Language (SQL)
Mark Smith
Purdue University
Supply Chain Management Technologies
Lee Sproull
New York University
Online Communities
Charles Steinfield
Michigan State University
Click-and-Brick Electronic Commerce
Electronic Commerce and Electronic Business
Edward A. Stohr
Stevens Institute of Technology
Online Analytical Processing (OLAP)
Dianna L. Stone
University of Central Florida
Human Resources Management
David Stotts
University of North Carolina at Chapel Hill
Perl
Judy Strauss
University of Nevada, Reno
Marketing Communication Strategies
Wayne C. Summers
Columbus State University
Local Area Networks
Jamie S. Switzer
Colorado State University
Virtual Teams
Dale R. Thompson
University of Arkansas
Public Networks
John S. Thompson
University of Colorado at Boulder
Integrated Services Digital Network (ISDN): Narrowband and Broadband Services and Applications
Stephen W. Thorpe
Neumann College
Extranets
Ronald R. Tidd
Central Washington University
Knowledge Management
Herbert Tuttle
The University of Kansas
Video Streaming
Okechukwu C. Ugweje
The University of Akron
Radio Frequency and Wireless Communications
Asoo J. Vakharia
University of Florida
Supply Chain Management
Robert Vaughn
University of Memphis
Law Enforcement
Vasja Vehovar
University of Ljubljana, Slovenia
Benchmarking Internet
Kim-Phuong L. Vu
Purdue University
Human Factors and Ergonomics
Jordan Walters
BCN Associates, Inc.
Managing a Network Environment
Siaw-Peng Wan
Elmhurst College
Online Banking and Beyond: Internet-Related Offerings from U.S. Banks
Youcheng Wang
University of Illinois at Urbana–Champaign
Travel and Tourism
James. L. Wayman
San Jose State University
Biometric Authentication
Scott Webster
Syracuse University
International Supply Chain Management
Jianbin Wei
Wayne State University
Load Balancing on the Internet
Ralph D. Westfall
California State Polytechnic University, Pomona
Telecommuting and Telework

Pamela Whitehouse
Harvard University
Distance Learning (Virtual Learning)
Dave Whitmore
Champlain College
Multiplexing
Russell S. Winer
New York University
Customer Relationship Management on the Web
Raymond Wisman
Indiana University Southeast
Web Search Fundamentals
Paul L. Witt
University of Texas at Arlington
Internet Relay Chat (IRC)
Mary Finley Wolfinbarger
California State University, Long Beach
Consumer Behavior
Peter R. Wurman
North Carolina State University
Online Auction Site Management
Cheng-Zhong Xu
Wayne State University
Load Balancing on the Internet
Qiang Yang
Hong Kong University of Science and Technology, China
Machine Learning and Data Mining on the Web
A. Neil Yerkey
University at Buffalo
Databases on the Web
Clement Yu
University of Illinois at Chicago
Web Search Technology
Daniel Dajun Zeng
University of Arizona
Intelligent Agents
Yan-Qing Zhang
Georgia State University
Fuzzy Logic
Xiaobo Zhou
University of Colorado at Colorado Springs
Load Balancing on the Internet
Donald E. Zimmerman
Colorado State University
Usability Testing: An Evaluation Process for Internet Communications

Preface

The Internet Encyclopedia is the first comprehensive examination of the core topics in the Internet field. *The Internet Encyclopedia*, a three-volume reference work with 205 chapters and more than 2,600 pages, provides comprehensive coverage of the Internet as a business tool, IT platform, and communications and commerce medium. The audience includes the libraries of two-year and four-year colleges and universities with MIS, IT, IS, data processing, computer science, and business departments; public and private libraries; and corporate libraries throughout the world. It is the only comprehensive source for reference material for educators and practitioners in the Internet field.

Education, libraries, health, medical, biotechnology, military, law enforcement, accounting, law, justice, manufacturing, financial services, insurance, communications, transportation, aerospace, energy, and utilities are among the fields and industries expected to become increasingly dependent upon the Internet and Web technologies. Companies in these areas are actively researching the many issues surrounding the design, utilization, and implementation of these technologies.

This definitive three-volume encyclopedia offers coverage of both established and cutting-edge theories and developments of the Internet as a technical tool and business/communications medium. The encyclopedia contains chapters from global experts in academia and industry. It offers the following unique features:

1) Each chapter follows a format which includes title and author, chapter outline, introduction, body, conclusion, glossary, cross references, and references. This unique format enables the readers to pick and choose among various sections of a chapter. It also creates consistency throughout the entire series.
2) The encyclopedia has been written by more than 240 experts and reviewed by more than 840 academics and practitioners chosen from around the world. This diverse collection of expertise has created the most definitive coverage of established and cutting edge theories and applications in this fast-growing field.
3) Each chapter has been rigorously peer reviewed. This review process assures the accuracy and completeness of each topic.
4) Each chapter provides extensive online and offline references for additional readings. This will enable readers to further enrich their understanding of a given topic.
5) More than 1,000 illustrations and tables throughout the series highlight complex topics and assist further understanding.
6) Each chapter provides extensive cross references. This helps the readers identify other chapters within the encyclopedia related to a particular topic, which provides a one-stop knowledge base for a given topic.
7) More than 2,500 glossary items define new terms and buzzwords throughout the series, which assists readers in understanding concepts and applications.
8) The encyclopedia includes a complete table of contents and index sections for easy access to various parts of the series.
9) The series emphasizes both technical and managerial issues. This approach provides researchers, educators, students, and practitioners with a balanced understanding of the topics and the necessary background to deal with problems related to Internet-based systems design, implementation, utilization, and management.
10) The series has been designed based on the current core course materials in several leading universities around the world and current practices in leading computer- and Internet-related corporations. This format should appeal to a diverse group of educators, practitioners, and researchers in the Internet field.

We chose to concentrate on fields and supporting technologies that have widespread applications in the academic and business worlds. To develop this encyclopedia, we carefully reviewed current academic research in the Internet field at leading universities and research institutions around the world. Management information systems, decision support systems (DSS), supply chain management, electronic commence, network design and management, and computer information systems (CIS) curricula recommended by the Association of Information Technology Professionals (AITP) and the Association for Computing Management (ACM) were carefully investigated. We also researched the current practices in the Internet field used by leading IT corporations. Our work enabled us to define the boundaries and contents of this project.

TOPIC CATEGORIES

Based on our research we identified 11 major topic areas for the encyclopedia:

- Foundation;
- Infrastructure;
- Legal, social, organizational, international, and taxation issues;
- Security issues and measures;
- Web design and programming;
- Design, implementation, and management;
- Electronic commerce;
- Marketing and advertising on the Web;

- Supply chain management;
- Wireless Internet and e-commerce; and
- Applications.

Although these 11 categories of topics are interrelated, each addresses one major dimension of the Internet-related fields. The chapters in each category are also interrelated and complementary, enabling readers to compare, contrast, and draw conclusions that might not otherwise be possible.

Although the entries have been arranged alphabetically, the light they shed knows no bounds. The encyclopedia provides unmatched coverage of fundamental topics and issues for successful design, implementation, and utilization of Internet-based systems. Its chapters can serve as material for a wide spectrum of courses, such as the following:

- Web technology fundamentals;
- E-commerce;
- Security issues and measures for computers, networks, and online transactions;
- Legal, social, organizational, and taxation issues raised by the Internet and Web technology;
- Wireless Internet and e-commerce;
- Supply chain management;
- Web design and programming;
- Marketing and advertising on the Web; and
- The Internet and electronic commerce applications.

Successful design, implementation, and utilization of Internet-based systems require a thorough knowledge of several technologies, theories, and supporting disciplines. Internet and Web technologies researchers and practitioners have had to consult many resources to find answers. Some of these sources concentrate on technologies and infrastructures, some on social and legal issues, and some on applications of Internet-based systems. This encyclopedia provides all of this relevant information in a comprehensive three-volume set with a lively format.

Each volume incorporates core Internet topics, practical applications, and coverage of the emerging issues in the Internet and Web technologies field. Written by scholars and practitioners from around the world, the chapters fall into the 11 major subject areas mentioned previously.

Foundation

Chapters in this group examine a broad range of topics. Theories and concepts that have a direct or indirect effect on the understanding, role, and the impact of the Internet in public and private organizations are presented. They also highlight some of the current issues in the Internet field. These articles explore historical issues and basic concepts as well as economic and value chain concepts. They address fundamentals of Web-based systems as well as Web search issues and technologies. As a group they provide a solid foundation for the study of the Internet and Web-based systems.

Infrastructure

Chapters in this group explore the hardware, software, operating systems, standards, protocols, network systems, and technologies used for design and implementation of the Internet and Web-based systems. Thorough discussions of TCP/IP, compression technologies, and various types of networks systems including LANs, MANS, and WANs are presented.

Legal, Social, Organizational, International, and Taxation Issues

These chapters look at important issues (positive and negative) in the Internet field. The coverage includes copyright, patent and trademark laws, privacy and ethical issues, and various types of cyberthreats from hackers and computer criminals. They also investigate international and taxation issues, organizational issues, and social issues of the Internet and Web-based systems.

Security Issues and Measures

Chapters in this group provide a comprehensive discussion of security issues, threats, and measures for computers, network systems, and online transactions. These chapters collectively identify major vulnerabilities and then provide suggestions and solutions that could significantly enhance the security of computer networks and online transactions.

Web Design and Programming

The chapters in this group review major programming languages, concepts, and techniques used for designing programs, Web sites, and virtual storefronts in the e-commerce environment. They also discuss tools and techniques for Web content management.

Design, Implementation, and Management

The chapters in this group address a host of issues, concepts, theories and techniques that are used for design, implementation, and management of the Internet and Web-based systems. These chapters address conceptual issues, fundamentals, and cost benefits and returns on investment for Internet and e-business projects. They also present project management and control tools and techniques for the management of Internet and Web-based systems.

Electronic Commerce

These chapters present a thorough discussion of electronic commerce fundamentals, taxonomies, and applications. They also discuss supporting technologies and applications of e-commerce inclining intranets, extranets, online auctions, and Web services. These chapters clearly demonstrate the successful applications of the Internet and Web technologies in private and public sectors.

Marketing and Advertising on the Web

The chapters in this group explore concepts, theories, and technologies used for effective marketing and advertising

on the Web. These chapters examine both qualitative and quantitative techniques. They also investigate the emerging technologies for mass personalization and customization in the Web environment.

Supply Chain Management

The chapters in this group discuss the fundamentals concepts and theories of value chain and supply chain management. The chapters examine the major role that the Internet and Web technologies play in an efficient and effective supply chain management program.

Wireless Internet and E-commerce

These chapters look at the fundamental concepts and technologies of wireless networks and wireless computing as they relate to the Internet and e-commerce operations. They also discuss mobile commerce and wireless marketing as two of the growing fields within the e-commerce environment.

Applications

The Internet and Web-based systems are everywhere. In most cases they have improved the efficiency and effectiveness of managers and decision makers. Chapters in this group highlight applications of the Internet in several fields, such as accounting, manufacturing, education, and human resources management, and their unique applications in a broad section of the service industries including law, law enforcement, medical delivery, health insurance and managed care, library management, nonprofit organizations, banking, online communities, dispute resolution, news services, public relations, publishing, religion, politics, and real estate. Although these disciplines are different in scope, they all utilize the Internet to improve productivity and in many cases to increase customer service in a dynamic business environment.

Specialists have written the collection for experienced and not-so-experienced readers. It is to these contributors that I am especially grateful. This remarkable collection of scholars and practitioners has distilled their knowledge into a fascinating and enlightening one-stop knowledge base in Internet-based systems that "talk" to readers. This has been a massive effort but one of the most rewarding experiences I have ever undertaken. So many people have played a role that it is difficult to know where to begin.

I should like to thank the members of the editorial board for participating in the project and for their expert advice on the selection of topics, recommendations for authors, and review of the materials. Many thanks to the more than 840 reviewers who devoted their times by proving advice to me and the authors on improving the coverage, accuracy, and comprehensiveness of these materials.

I thank my senior editor at John Wiley & Sons, Matthew Holt, who initiated the idea of the encyclopedia back in spring of 2001. Through a dozen drafts and many reviews, the project got off the ground and then was managed flawlessly by Matthew and his professional team. Matthew and his team made many recommendations for keeping the project focused and maintaining its lively coverage. Tamara Hummel, our superb editorial coordinator, exchanged several hundred e-mail messages with me and many of our authors to keep the project on schedule. I am grateful to all her support. When it came to the production phase, the superb Wiley production team took over. Particularly I want to thank Deborah DeBlasi, our senior production editor at John Wiley & Sons, and Nancy J. Hulan, our project manager at TechBooks. I am grateful to all their hard work.

Last, but not least, I want to thank my wonderful wife Nooshin and my two lovely children Mohsen and Morvareed for being so patient during this venture. They provided a pleasant environment that expedited the completion of this project. Nooshin was also a great help in designing and maintaining the author and reviewer databases. Her efforts are greatly appreciated. Also, my two sisters Azam and Akram provided moral support throughout my life. To this family, any expression of thanks is insufficient.

Hossein Bidgoli
California State University, Bakersfield

Guide to the Internet Encyclopedia

The Internet Encyclopedia is a comprehensive summary of the relatively new and very important field of the Internet. This reference work consists of three separate volumes and 205 chapters on various aspects of this field. Each chapter in the encyclopedia provides a comprehensive overview of the selected topic intended to inform a board spectrum of readers ranging from computer professionals and academicians to students to the general business community.

In order that you, the reader, will derive the greatest possible benefit from *The Internet Encyclopedia,* we have provided this Guide. It explains how the information within the encyclopedia can be located.

ORGANIZATION

The Internet Encyclopedia is organized to provide maximum ease of use for its readers. All of the chapters are arranged in alphabetical sequence by title. Chapters titles that begin with the letters A to F are in Volume 1, chapter titles from G to O are in Volume 2, and chapter titles from P to Z are in Volume 3. So that they can be easily located, chapter titles generally begin with the key word or phrase indicating the topic, with any descriptive terms following. For example, "Virtual Reality on the Internet: Collaborative Virtual Reality" is the chapter title rather than "Collaborative Virtual Reality."

Table of Contents

A complete table of contents for the entire encyclopedia appears in the front of each volume. This list of titles represents topics that have been carefully selected by the editor-in-chief, Dr. Hossein Bidgoli, and his colleagues on the Editorial Board.

Following this list of chapters by title is a second complete list, in which the chapters are grouped according to subject area. The encyclopedia provides coverage of 11 specific subject areas, such as E-commerce and Supply Chain Management. Please see the Preface for a more detailed description of these subject areas.

Index

The Subject Index is located at the end of Volume 3. This index is the most convenient way to locate a desired topic within the encyclopedia. The subjects in the index are listed alphabetically and indicate the volume and page number where information on this topic can be found.

Chapters

Each chapter in *The Internet Encyclopedia* begins on a new page, so that the reader may quickly locate it. The author's name and affiliation are displayed at the beginning of the article.

All chapters in the encyclopedia are organized according to a standard format, as follows:

- Title and author,
- Outline,
- Introduction,
- Body,
- Conclusion,
- Glossary,
- Cross References, and
- References.

Outline

Each chapter begins with an outline indicating the content to come. This outline provides a brief overview of the chapter so that the reader can get a sense of the information contained there without having to leaf through the pages. It also serves to highlight important subtopics that will be discussed within the chapter. For example, the chapter "Computer Literacy" includes sections entitled Defining a Computer, Categories of Computers According to Their Power, and Classes of Data Processing Systems. The outline is intended as an overview and thus lists only the major headings of the chapter. In addition, lower-level headings will be found within the chapter.

Introduction

The text of each chapter begins with an introductory section that defines the topic under discussion and summarizes the content. By reading this section the readers get a general idea about the content of a specific chapter.

Body

The body of each chapter discusses the items that were listed in the outline section.

Conclusion

The conclusion section provides a summary of the materials discussed in each chapter. This section imparts to the readers the most important issues and concepts discussed within each chapter.

Glossary

The glossary contains terms that are important to an understanding of the chapter and that may be unfamiliar to the reader. Each term is defined in the context of the particular chapter in which it is used. Thus the same term may be defined in two or more chapters with the detail of the definition varying slightly from one to another. The encyclopedia includes approximately 2,500 glossary terms.

For example, the article "Computer Literacy" includes the following glossary entries:

Computer A machine that accepts data as input, processes the data without human interference using a set of stored instructions, and outputs information. Instructions are step-by-step directions given to a computer for performing specific tasks.

Computer generations Different classes of computer technology identified by a distinct architecture and technology; the first generation was vacuum tubes, the second transistors, the third integrated circuits, the fourth very-large-scale integration, and the fifth gallium arsenide and parallel processing.

Cross References

All the chapters in the encyclopedia have cross references to other chapters. These appear at the end of the chapter, following the text and preceding the references. The cross references indicate related chapters which can be consulted for further information on the same topic. The encyclopedia contains more than 2,000 cross references in all. For example, the chapter "Java" has the following cross references:

JavaBeans and Software Architecture; Software Design and Implementation in the Web Environment.

References

The reference section appears as the last element in a chapter. It lists recent secondary sources to aid the reader in locating more detailed or technical information. Review articles and research papers that are important to an understanding of the topic are also listed. The references in this encyclopedia are for the benefit of the reader, to provide direction for further research on the given topic. Thus they typically consist of one to two dozen entries. They are not intended to represent a complete listing of all materials consulted by the author in preparing the chapter. In addition, some chapters contain a Further Reading section, which includes additional sources readers may wish to consult.

P

Passwords

Jeremy Rasmussen, *Sypris Electronics, LLC*

INTRODUCTION

The ancient folk tale of Ali Baba and the forty thieves mentions the use of a password. In this story, Ali Baba finds that the phrase "Open Sesame" magically opens the entrance to a cave where the thieves have hidden their treasure. Similarly, modern computer systems use passwords to authenticate users and allow them entrance to system resources and data shares on an automated basis. The use of passwords in computer systems likely can be traced to the earliest timesharing and dial-up networks. Passwords were probably not used before then in purely batch systems.

The security provided by a password system depends on the passwords being kept secret at all times. Thus, a password is vulnerable to compromise whenever it is used, stored, or even known. In a password-based authentication mechanism implemented on a computer system, passwords are vulnerable to compromise due to five essential aspects of the password system:

Passwords must be initially assigned to users when they are enrolled on the system;
Users' passwords must be changed periodically;
The system must maintain a "password database";
Users must remember their passwords; and
Users must enter their passwords into the system at authentication time.

Because of these factors, a number of protection schemes have been developed for maintaining password security. These include implementing policies and mechanisms to ensure "strong" passwords, encrypting the password database, and simplifying the sign-on and password synchronization processes. Even so, a number of sophisticated cracking tools are available today that threaten password security. For that reason, it is often advised that passwords be combined with some other form of security to achieve strong authentication.

TYPES OF IDENTIFICATION/ AUTHENTICATION

Access control is the security service that deals with granting or denying permission for subjects (e.g., users or programs) to use objects (e.g., other programs or files) on a given computer system. Access control can be accomplished through either hardware or software features, operating procedures, management procedures, or a combination of these. Access control mechanisms are classified by their ability to verify the authenticity of a user. The three basic verification methods are as follows:

What you have (examples: smart card or token);
What you are (examples: biometric fingerprint [see Figure 1] or iris pattern); and
What you know (examples: PIN or password).

Of all verification methods, passwords are probably weakest, yet they are still the most widely used method in systems today. In order to guarantee strong

Figure 1: A biometric fingerprint scanner.

authentication, a system ought to combine two or more of these factors. For example, in order to access an ATM, one must have a bank card and know his or her personal identification number (PIN).

HISTORY OF PASSWORDS IN MODERN COMPUTING

Conjecture as to which system was the first to incorporate passwords has been bandied about by several computing pioneers on the Cyberspace History List-Server (CYHIST). However, there has not been any concrete evidence as yet to support one system or another as the progenitor. The consensus opinion favors the Compatible Time Sharing System (CTSS) developed at the Massachusetts Institute of Technology (MIT) Computation Center beginning in 1961. As part of Project MAC (Multiple Access Computer) under the direction of Professor Fernando J. "Corby" Corbató, the system was implemented on an IBM 7094 and reportedly began using passwords by 1963. According to researcher Norman Hardy, who worked on the project, the security of passwords immediately became an issue as well: "I can vouch for some version of CTSS having passwords. It was in the second edition of the CTSS manual, I think, that illustrated the login command. It had Corby's user name and password. It worked—and he changed it the same day."

Passwords were widely in use by the early 1970s as the "hacker" culture began to develop, possibly in tacit opposition to the ARPANET. Now, with the explosion of the Internet, the use of passwords and the quantity of confidential data that those passwords protect have grown exponentially. But just as the 40 thieves' password protection system was breached (the cave could not differentiate between Ali Baba's voice and those of the thieves), computer password systems have also been plagued by a number of vulnerabilities. Although strong password authentication has remained a "hard" problem in cryptography despite advances in both symmetric (secret-key) and asymmetric (public-key) cryptosystems, the history of password authentication is replete with examples of weak, easily compromised systems. In general, "weak" authentication systems are characterized by protocols that either leak the password directly over the network or leak sufficient information while performing authentication to allow intruders to deduce or guess at the password.

Green Book: The Need for Accountability

In 1983, the U.S. Department of Defense Computer Security Center (CSC) published the venerable tome *Trusted Computer System Evaluation Criteria,* also known as the Orange Book. This publication defined the assurance requirements for security protection of computer systems that were to be used in processing classified or other sensitive information. One major requirement imposed by the Orange Book was accountability: "Individual accountability is the key to securing and controlling any system that processes information on behalf of individuals or groups of individuals" (Latham, 1985).

The Orange Book clarified accountability as follows:

Individual user identification: Without this, there is no way to distinguish the actions of one user on a system from those of another.

Authentication: Without this, user identification has no credibility. And without a credible identity, no security policies can be properly invoked because there is no assurance that proper authorizations can be made.

The CSC went on to publish the *Password Management Guideline* (also known as the Green Book) in 1985 "to assist in providing that much needed credibility of user identity by presenting a set of good practices related to the design, implementation and use of password-based user authentication mechanisms." The Green Book outlined a number of steps that system security administrators should take to ensure password security on the system and suggests that, whenever possible, they be automated. These include the following 10 rules (Brotzman, 1985):

System security administrators should change the passwords for all standard user IDs before allowing the general user population to access the system.

A new user should always appear to the system as having an "expired password" which will require the user to change the password by the usual procedure before receiving authorization to access the system.

Each user ID should be assigned to only one person. No two people should ever have the same user ID at the same time, or even at different times. It should be considered a security violation when two or more people know the password for a user ID.

Users need to be aware of their responsibility to keep passwords private and to report changes in their user status, suspected security violations, etc. Users should also be required to sign a statement to acknowledge understanding of these responsibilities.

Passwords should be changed on a periodic basis to counter the possibility of undetected password compromise.

Users should memorize their passwords and not write them on any medium. If passwords must be written, they should be protected in a manner that is consistent with the damage that could be caused by their compromise.

Stored passwords should be protected by access controls provided by the system, by password encryption, or by both.

Passwords should be encrypted immediately after entry, and the memory containing the plaintext password should be erased immediately after encryption.

Only the encrypted password should be used in comparisons. There is no need to be able to decrypt passwords. Comparisons can be made by encrypting the password entered at login and comparing the encrypted form with the encrypted password stored in the password database.

The system should not echo passwords that users type in, or at least should mask the entered password (e.g., with asterisks).

PASSWORD SECURITY—BACKGROUND

Information Theory

Cryptography is a powerful mechanism for securing data and keeping them confidential. The idea is that the original message is scrambled via an algorithm (or cipher), and only those with the correct key can unlock the scrambled message and get back the plaintext contents. In general, the strength of a cryptographic algorithm is based on the length and quality of its keys. Passwords are a similar problem. Based on their length and quality, they should be more difficult to attack either by dictionary, by hybrid, or by brute-force attacks. However, the quality of a password, just as the quality of a cryptographic key, is based on entropy. Entropy is a measure of disorder.

An example of entropy

Say a user is filling out a form on a Web page (see Figure 2). The form has a space for "Sex," and leaves six characters for entering either "female" or "male" before encrypting the form entry and sending it to the server. If each character is a byte (i.e., 8 bits), then $6 \times 8 = 48$ bits will be sent for this response. Is this how much information is actually contained in the field, though?

Clearly, there is only one bit of data represented by the entry—a binary value—either male or female. That means that there is only one bit of entropy (or uncertainty) and there are 47 bits of redundancy in the field. This redundancy could be used by a cryptanalyst (someone who analyzes cryptosystems) to help crack the key.

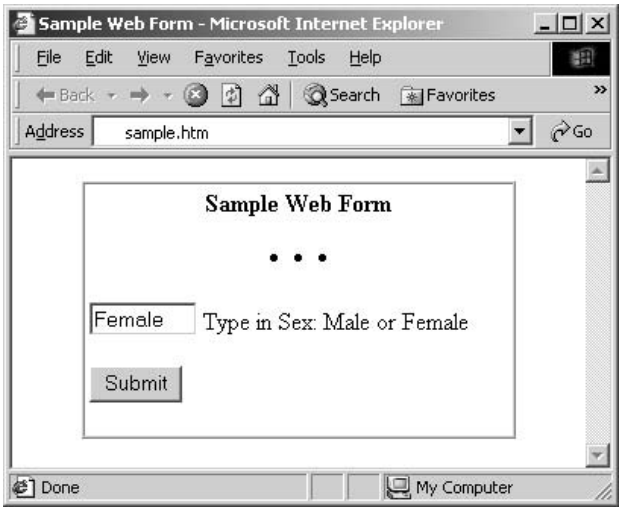


Figure 2: Sample Web page entry form.

Fundamental work by Claude Shannon during the 1940s illustrated this concept, that is, that the amount of information in a message is not necessarily a function of the length of a message (or the number of symbols used in the message) (Sloane & Wyner, 1993). Instead, the amount of information in a message is determined by how many different possible messages there are and how frequently each message is used.

The same concepts apply to password security. A longer password is not necessarily a better password. Rather, a password that is difficult to guess (i.e., one that has high entropy) is best. This usually comes from a combination of factors (see "Guidelines for selecting a good password"). The probability that any single attempt at guessing a password will be successful is one of the most critical factors in a password system. This probability depends on the size of the password space and the statistical distribution within that space of passwords that are actually used.

Over the past several decades, Moore's Law has made it possible to brute-force password spaces of larger and larger entropy. In addition, there is a limit to the entropy that the average user can remember. A user cannot typically remember a 32-character password, but that is what is required to have the equivalent strength of a 128-bit key. Recently, password cracking tools have advanced to the point of being able to crack nearly anything a system could reasonably expect a user to memorize (see "Password Length and Human Memory").

Cryptographic Protection of Passwords

Early on, the most basic and least secure method of authentication was to store passwords in plaintext (i.e., unencrypted) in a database on the server. During authentication, the client would send his or her password to the server, and the server would compare this against the stored value. Obviously, however, if the password file were accessible to unauthorized users, the security of the system could be easily compromised.

In later systems, developers discovered that a server did not have to store a user's password in plaintext form in order to perform password authentication. Instead, the user's password could be transformed through a one-way function, such as a hashing function, into a random-looking sequence of bytes. Such a function would be difficult to invert. In other words, given a password, it would be easy to compute its hash, but given a hash, it would be computationally infeasible to compute the password from it (see "Hashing"). Authentication would consist merely of performing the hash function over the client's password and comparing it to the stored value. The password database itself could be made accessible to all users without fear of an intruder being able to steal passwords from it.

Hashing

A hash function is an algorithm that takes a variable-length string as the input and produces a fixed-length value (hash) as the output. The challenge for a hashing algorithm is to make this process irreversible; that is, finding

Table 1 Output from the MD5 Test Suite

For the Input String	The Output Message Digest is
"" (no password)	d41d8cd98f00b204e9800998ecf8427e
"a"	0cc175b9c0f1b6a831c399e269772661
"abc"	900150983cd24fb0d6963f7d28e17f72
"message digest"	f96b697d7cb7938d525a2f31aaf161d0
"abcdefghijklmnopqrstuvwxyz"	c3fcd3d76192e4007dfb496cca67e13b
"ABCDEFGHIJKLMNOPQRSTUVWXYZabcdefghijklmno pqrstuvwxyz0123456789"	d174ab98d277d9f5a5611c2c9f419d9f
"12345678901234567890123456789012345678901234567890 12345678901234567890123456789012345678901234567890"	57edf4a22be3c955ac49da2e2107b67a

a string that produces a given hash value should be very difficult. It should also be difficult to find two arbitrary strings that produce the same hash value. Also called a message digest or fingerprint, several one-way hash functions are in common use today. Among these are Secure Hashing Algorithm-1 (SHA-1) and Message Digest-5 (MD-5). The latter was invented by Ron Rivest for RSA Security, Inc. and produces a 128-bit hash value. See Table 1 for an example of output generated by MD5. SHA-1 was developed by the U.S. National Institute of Standards and Technology (NIST) and the National Security Agency (NSA) and produces 160-bit hash values. SHA-1 is generally considered more secure than MD5 due to its longer hash value.

Microsoft Windows NT uses one-way hash functions to store password information in the Security Account Manager (SAM). There are no Windows32 Applications Programming Interface (API) function calls to retrieve user passwords because the system does not store them. It stores only hash values. However, even a hash-encrypted password in a database is not entirely secure. A cracking tool can compile a list of, say, the one million most commonly used passwords and compute hash functions from all of them. Then the tool can obtain the system account database and compare the hashed passwords in the database with its own list to see what matches. This is called a "dictionary attack" (see "Password Cracking Tools").

To make dictionary attacks more difficult, often a salt is used. A salt is a random string that is concatenated with a password before it is operated on by the hashing function. The salt value is then stored in the user database, together with the result of the hash function. Using a salt makes dictionary attacks more difficult, as a cracker would have to compute the hashes for all possible salt values.

A simple example of a salt would be to add the time of day; for example, if a user logs in at noon using the password "pass," the string that would be encrypted might be "1p2a0s0s." By adding this randomness to the password, the hash will actually be different every time the user logs in (unless it is at noon every day). Whether a salt is used and what the salt actually is depends upon the operating system and the encryption algorithm being used. On a FreeBSD system, for example, there is a function called crypt that uses the DES, MD5, or Blowfish algorithms to hash passwords and can also use three forms of salts.

According to Cambridge University professor of computing Roger Needham, the Cambridge Multiple Access System (CMAS), which was an integrated online–offline terminal or regular input-driven system, may have been among the earliest to implement such one-way functions. It first went online in 1967 and incorporated password protection. According to Needham: "In 1966, we conceived the use of one-way functions to protect the password file, and this was an implemented feature from day one" (R. Needham, personal communication, April 11, 2002).

One-way hashing is still being used today, although it does not address another weakness—in a networked environment, it is difficult to transmit the password securely to the server for verification without its being captured and reused, perhaps in a replay attack. To avoid revealing passwords directly over an untrusted network, computer scientists have developed challenge–response systems. At their simplest, the server sends the user some sort of challenge, which would typically be a random string of characters called a nonce. The user then computes a response, usually some function based on both the challenge and the password. This way, even if the intruder captured a valid challenge–response pair, it would not help him or her gain access to the system, because future challenges would be different and require different responses.

These challenge-and-response systems are referred to as one-time password (OTP) systems. Bellcore's S/KEY is one such system in which a one-time password is calculated by combining a seed with a secret password known only to the user and then applying a secure hashing algorithm a number of times equal to the sequence number. Each time the user is authenticated, the sequence number expected by the system is decremented, thus eliminating the possibility of an attacker trying a replay attack using the same password again. One-time passwords were more prevalent before secure shell (SSH) and secure sockets layer (SSL) systems came into widespread use.

PASSWORD CRACKING TOOLS

Password-Cracking Approaches

As mentioned earlier, passwords are typically stored as values hashed with SHA-1 or MD5, which are one-way functions. In other words, this entire encyclopedia could be hashed and represented as eight bytes of gibberish. There would be no way to use these eight bytes of data to obtain the original text. However, password crackers know that people do not use whole encyclopedias as their passwords. The vast majority of passwords are 4 to

12 characters. Passwords are also, in general, not just random strings of symbols. Because users need to remember them, passwords are usually words or phrases of significance to the user. This is an opportunity for the attacker to reduce the search space.

An attacker might steal a password file–or sniff the wire and capture the user ID/password hash pairs during logon–and then run a password-cracking tool on it. Because it is impossible to decrypt a hash back to a password, these programs will try a dictionary approach first. The program guesses a password—say, the word "Dilbert." The program then hashes "Dilbert" and compares the hash to one of the hashed entries in the password file. If it matches, then that password hash represents the password "Dilbert." If the hash does not match, the program takes another guess. Depending on the tool, a password cracker will try all the words in a dictionary, all the names in a phone book, and so on. Again, the attacker does not need to know the original password–just a password that hashes to the same value.

This is analogous to the "birthday paradox," which basically says, "If you get 25 people together in a room, the odds are better than 50/50 that two of them will have the same birthday." How does this work? Imagine a person meeting another on the street and asking him his birthday. The chances of the two having the same birthday are only 1/365 (0.27%). Even if one person asks 25 people, the probability is still low. But with 25 people in a room together, each of the 25 is asking the other 24 about their birthdays. Each person only has a small (less than 5%) chance of success, but trying it 25 times increases the probability significantly.

In a room of 25 people, there are 300 possible pairs (25*24/2). Each pair has a probability of success of $1/365 = 0.27\%$, and a probability of failure of $1 - 0.27\% = 99.726\%$. Calculating the probability of failure: $99.726\%^{300} = 44\%$. The probability of success is then $100\% - 44\% = 56\%$. So a birthday match will actually be found five out of nine times. In a room with 42 people, the odds of finding a birthday match rise to 9 out of 10. Thus, the birthday paradox is that it is much easier to find two values that match than it is to find a match to some particular value.

If a wave of dictionary guesses fails to produce any passwords for the attacker, the cracking program will next try a hybrid approach of different combinations—-such as forward and backward spellings of dictionary words, additional numbers and special characters, or sequences of characters. The goal here again is to reduce the cracker's search space by trying "likely" combinations of known words.

Only after exhausting both of these avenues will the cracking program start in on an exhaustive or brute-force attack on the entire password space. And, of course, it remembers the passwords it has already tried and will not have to recheck these either during the brute-force search.

Approaches to Retrieving Passwords

Most password-cracking programs will first attempt to retrieve password hashes to begin their cracking processes. A sophisticated attacker will not try to guess passwords by entering them through the standard user interface because the time to do so is prohibitive, and most systems can be configured to lock a user out after too many wrong guesses.

On Microsoft Windows systems, it typically requires the "Administrator" privilege to read the password hashes from the database in which they are stored. This is usually somewhere in the system registry. In order to access them, a cracking tool will attempt to dump the password hashes from the Windows registry on the local machine or over the network if the remote machine allows network registry access. The latter requires a target Windows machine name or IP address.

Another method is to access the password hashes directly from the file system. On Microsoft Windows systems, this is the SAM. Because Windows locks the SAM file where the password hashes are stored in the file system with an encryption mechanism known as SYSKEY, it is impossible to read them from this file while the system is running. However, sometimes there is a backup of this file on tape, on an emergency repair disk (ERD), or in the repair directory of the system's hard drive. Alternately, a user may boot from a floppy disk running another operating system such as MS-DOS and be able to read password hashes directly from the file system. This is why security administrators should never neglect physical security of systems. If an attacker can physically access a machine, he or she can bypass the built-in file system security mechanisms (see Recovering Windows NT Passwords).

Todd Sabin has released a free utility called PWDUMP2 that can dump the password hashes on a local machine if the SAM has been encrypted with the SYSKEY utility that was introduced in Windows NT Service Pack 3. Once a user downloads the utility, he or she can follow the instructions on the Web page to retrieve the password hashes, load the hashes into a tool such as L0phtCrack, and begin cracking them.

Password Sniffing

Instead of capturing the system user file (SAM on Windows or /etc/passwd or /etc/shadow on Unix/Linux), another way of collecting user IDs and passwords is through sniffing network traffic. Sniffing uses some sort of software or hardware wiretap device to eavesdrop on network communications, usually by capturing and deciphering communications packets. According to Peiter "Mudge" Zatko, who initially wrote L0phtCrack: "Sniffing is slang for placing a network card into promiscuous mode so that it actually looks at all of the traffic coming along the line and not just the packets that are addressed to it. By doing this one can catch passwords, login names, confidential information, etc" (Zatko, 1999b).

L0phtCrack offers an "SMB Packet Capture" function to capture encrypted hashes transmitted over a Windows network segment. On a switched network, a cracker will only be able to sniff sessions originating from the local machine or connecting to that machine. As server message block (SMB) session authentication messages are captured by the tool, they are displayed in the SMB Packet Capture window. The display shows the source and destination IP addresses, the user name, the SMB challenge, the encrypted LAN manager hash, and the encrypted NT LAN manager hash, if any. To crack these hashes,

the tool saves the session and then works on the captured file.

Recovering Windows NT Passwords

Or, why physical security is still important. Norwegian software developer Petter Nordahl-Hagen has built a resource ("The Offline NT Password Editor") for recovering Windows passwords on workstations. His approach bypasses the NTFS file permissions of Windows NT, 2000, and XP by using a Linux boot disk that allows one to reset the Administrator password on a system by replacing the hash stored in the SAM with a user-selected hash. His program has even been shown to work on Windows 2000 systems with SYSKEY enabled. An MS-DOS version also exists, as does a version that boots from CD-ROM instead of floppy disk.

Thus, physical access to the workstation can mean instant compromise, unless, perhaps the system BIOS settings are also password-protected and do not allow a user to boot from floppy or CD-ROM (however, several attacks against BIOS settings have also been published).

Types of Password-Cracking Tools

Password-cracking tools can be divided into two categories—those that attempt to retrieve system-level login passwords and those that attack the password protection mechanisms of specific applications. The first type includes programs such as L0phtcrack, Cain & Abel, and John the Ripper. Some sites for obtaining password-cracking tools for various platforms, operating systems, and applications are included in the Further Reading section at the end of this chapter.

The Russian company ElcomSoft has a developed a range of programs that can crack passwords on Microsoft Office encrypted files, WinZip or PKZip archived files, or Adobe Acrobat (PDF) files. The U.S. federal government charged ElcomSoft with violating the Digital Millennium Copyright Act of 1998 for selling a program that allowed people to disable encryption software from Adobe Systems that is used to protect electronic books. The case drew attention after ElcomSoft programmer Dmitry Sklyarov was arrested at the DefCon 2001 convention in July, 2001 (US. ElcomSoft & Sklyarov FAQ, n.d.).

PASSWORD SECURITY ISSUES AND EFFECTIVE MANAGEMENT

Enforcing Password Guidelines

The FBI and the Systems Administration and Networking Security (SANS) Institute released a document summarizing the "Twenty Most Critical Internet Security Vulnerabilities." The majority of successful attacks on computer systems via the Internet can be traced to exploitation of security flaws on this list. One of items on this list is "accounts with no passwords or weak passwords." In general, these accounts should be removed or assigned stronger passwords. In addition, accounts with built-in or default passwords that have never been reconfigured create vulnerability because they usually have the same password across installations of the software. Attackers will look for these accounts, having found the commonly known passwords published on hacking Web sites or some other public forum. Therefore, any default or built-in accounts also need to be identified and removed from the system or else reconfigured with stronger passwords.

The list of common vulnerabilities and exposures (CVE) maintained by the MITRE Corporation (http://www.cve.mitre.org) provides a taxonomy for more than 2000 well-known attacker exploits. Among these, nearly 100 have to do with password insecurities, and another 250 having to do with passwords are "candidates" currently under review for inclusion in the list. The following provides a few samples:

Some Sample Password Vulnerabilities in the CVE List

CVE-1999–0366: "In some cases, Service Pack 4 for Windows NT 4.0 can allow access to network shares using a blank password, through a problem with a null NT hash value."

CVE-2001–0465: "TurboTax saves passwords in a temporary file when a user imports investment tax information from a financial institution, which could allow local users to obtain sensitive information."

CVE-2000–1187: "Buffer overflow in the HTML parser for Netscape 4.75 and earlier allows remote attackers to execute arbitrary commands via a long password value in a form field."

CVE-1999–1104: "Windows 95 uses weak encryption for the password list (.pwl) file used when password caching is enabled, which allows local users to gain privileges by decrypting the passwords."

CVE-2000–0981: "MySQL Database Engine uses a weak authentication method which leaks information that could be used by a remote attacker to recover the password."

CVE-2000–0267: "Cisco Catalyst 5.4.x allows a user to gain access to the 'enable' mode without a password."

CVE-1999–1298: "Sysinstall in FreeBSD 2.2.1 and earlier, when configuring anonymous FTP, creates the ftp user without a password and with/bin/date as the shell, which could allow attackers to gain access to certain system resources."

CVE-1999–1316: "Passfilt.dll in Windows NT SP2 allows users to create a password that contains the user's name, which could make it easier for an attacker to guess" (Common Vulnerabilities, n.d.).

SANS suggests that to determine if one's system is vulnerable to such attacks, one needs to be cognizant of all the user accounts on the system. First, the system security administrator must inventory the accounts on the system and create a master list. This list should include even intermediate systems, such as routers and gateways, as well as any Internet-connected printers and print controllers. Second, the administrator should develop procedures for adding authorized accounts to the list and for removing accounts when they are no longer in use. The master list should be validated on a regular basis. In addition, the administrator should run some password strength-checking tool against the accounts to look for weak or nonexistent passwords. A sample of these tools is noted in the Further Reading section at the end of this chapter.

Many organizations supplement password control programs with procedural or administrative controls that ensure that passwords are changed regularly and that old passwords are not reused. If password aging is used, the system should give users a warning and the opportunity to change their passwords before they expire. In addition, administrators should set account lockout policies, which lock out a user after a number of unsuccessful login attempts, and cause him or her to have his password reset.

Microsoft Windows 2000 and Windows XP include built-in password constraint options in the "Group Policy" settings. An administrator can configure the network so that user passwords must have a minimum length, a minimum and maximum age, and other constraints. It is important to require a minimum age on a password.

The following outlines the minimal criteria for selecting "strong" passwords.

Guidelines for Selecting a Good Password

The goal is to select something easily remembered but not easily guessed.

Length

Windows systems: seven characters or longer

Unix, Linux systems: eight characters or longer

Composition

Mixture of alphabetic, numeric, and special characters (e.g., #, @, or !)

Mixture of upper and lower case characters

No words found in a dictionary

No personal information about the user (e.g., any part of the user's name, a family member's name, or the user;s date of birth, Social Security number, phone number, license plate number, etc.)

No information that is easily obtained about the user, especially any part of the user ID

No commonly used proper names such as local sports teams or celebrities

No patterns such as 12345, sssss, or qwerty

Try misspelling or abbreviating a word that has some meaning to the user (Example: "How to select a good password?" becomes "H2sagP?")

Password Aging and Reuse

To limit the usefulness of passwords that might have been compromised, it is suggested practice to change them regularly. Many systems force users to change their passwords when they log in for the first time, and again if they have not changed their passwords for an extended period (say, 90 days). In addition, users should not reuse old passwords. Some systems support this by recording the old passwords, ensuring that users cannot change their passwords back to previously used values, and ensuring that the users' new passwords are significantly different from their previous passwords. Such systems usually have a finite memory, say the past 10 passwords, and users can circumvent the password filtering controls by changing a password 10 times in a row until it is the same as the previously used password.

It is recommended that, at a predetermined period of time prior to the expiration of a password's lifetime, the user ID it is associated with be notified by the system as having an "expired" password. A user who logs in with an ID having an expired password should be required to change the password for that user ID before further access to the system is permitted. If a password is not changed before the end of its maximum lifetime, it is recommended that the user ID it is associated with be identified by the system as "locked." No login should be permitted to a locked user ID, but the system administrator should be able to unlock the user ID by changing the password for that user ID. After a password has been changed, the lifetime period for the password should be reset to the maximum value established by the system.

Social Engineering

With all the advances in technology, the oldest way to attack a password-based security system is still the easiest: coercion, bribery, or trickery against the users of the system. Social engineering is an attack against people, rather than machines. It is an outsider's use of psychological tricks on legitimate users of a computer system, usually to gain the information (e.g., user IDs and passwords) needed to access a system. The notorious "hacker" Kevin Mitnick, who was convicted on charges of computer and wire fraud and spent 59 months in federal prison, told a Congressional panel that he rarely used technology to gain information and used social engineering almost exclusively (Federation of American Scientists, n.d.).

According to a study by British psychologists, people often base their passwords on something obvious and easily guessed by a social engineer. Around 50% of computer users base them on the name of a family member, a partner, or a pet. Another 30% use a pop idol or sporting hero. Another 10% of users pick passwords that reflect some kind of fantasy, often containing some sexual reference. The study showed that only 10% use cryptic combinations that follow all the rules of "tough" passwords (Brown, 2002).

The best countermeasures to social engineering attacks are education and awareness. Users should be instructed never to tell anyone their passwords. Doing so destroys accountability, and a system administrator should never need to know it either. Also, users should never write down their passwords. A clever social engineer will find it if it is "hidden" under a mouse pad or inside a desk drawer.

Some Examples of Social Engineering Attacks

"Appeal to Authority" Attack. This is impersonating an authority figure or else identifying a key individual as a supposed acquaintance, in order to demand information. For example: A secretary receives a phone call from someone claiming to be the "IT Manager." He requests her user ID and password, or gives her a value to set her password to immediately because "there has been a server crash in the computer center and we need to reset everyone's account." Once she has complied, he now has access to a valid user ID and password to access the system.

"Fake Web Site" Attack. The same password should not be used for multiple applications. Once a frequently used

password is compromised, all of the user's accounts will be compromised. A good social engineering attack might be to put up an attractive Web site with titillating content, requiring users to register a username and password in order to access the "free" information. The attacker would record all passwords (even incorrect ones, which a user might have mistakenly entered thinking of another account), and then use those to attack the other systems frequented by the user. The Web site could even solicit information from the users about their accounts—for example, what online brokerage, banking, and e-mail accounts they used. Web site operators can always keep a log of IP addresses used to access the site and could go back to attack the originating system directly.

"Dumpster Diving" Attack. Many serious compromises are still caused by contractors and third parties throwing away draft instruction manuals, development notes, etc., with user IDs and passwords in them. Social engineers may employ "dumpster diving," that is, digging through paper printouts in the trash looking for such significant information to gain system access.

Single Sign-On and Password Synchronization

One issue that has irritated users in large secure environments is the burgeoning number of passwords they have to remember to access various applications. A user might need one password to log onto his or her workstation, another to access the network, and yet another for a particular server. Ideally, a user should be able to sign on once, with a single password, and be able to access all the other systems on which he or she has authorization.

Some have called this notion of single sign-on the "Holy Grail" of computer security. The goal is admirable—to create a common enterprise security infrastructure to replace a heterogeneous one. And it is currently being attempted by several vendors through technologies such as the Open Group's Distributed Computing Environment (DCE), MIT's Kerberos, Microsoft's ActiveDirectory, and Public-Key Infrastructure (PKI)-based systems. However, few, if any, enterprises have actually achieved their goal. Unfortunately, the task of changing all existing applications to use a common security infrastructure is very difficult, and this has further been hampered by a lack of consensus on a common security infrastructure. As a result, the disparate proprietary and standards-based solutions cannot be applied to every system. In addition, there is a risk of a single point of failure. Should one user's password be compromised, it is not just his local system that can be breached but the entire enterprise.

Password synchronization is another means of trying to help users maintain the passwords that they use to log onto disparate systems. In this scheme, when users periodically change their passwords, the new password is applied to every account the user has, rather than just one. The main objective of password synchronization is to help users remember a single, strong password. Password synchronization purports to improve security because synchronized passwords are subjected to a strong password policy, and users who remember their passwords are less likely to write them down.

To mitigate the risk of a single system compromise being leveraged by an intruder into a network-wide attack:

Very insecure systems should not participate in a password synchronization system,

Synchronized passwords should be changed regularly, and

Users should be required to select strong (hard to guess) passwords when synchronization is introduced.

Unix/Linux-Specific Password Issues

Traditionally on Unix and Linux platforms, user information, including passwords, is kept in a system file called /etc/passwd. The password for each user is stored as a hash value. Despite the password being encoded with a one-way hash function and a salt as described earlier, a password cracker could still compromise system security if he or she obtained access to the /etc/passwd file and used a successful dictionary attack. This vulnerability can be mitigated by simply moving the passwords in the /etc/passwd file to another file, usually named /etc/shadow, and making this file readable only by those who have administrator or "root" access to the system.

In addition, Unix or Linux administrators should examine the password file (as well as the shadow password file when applicable) on a regular basis for potential account-level security problems. In particular, it should be examined for the following:

Accounts without passwords.

UIDs of 0 for accounts other than root (which are also superuser accounts).

GIDs of 0 for accounts other than root. Generally, users don't have group 0 as their primary group.

Other types of invalid or improperly formatted entries.

User names and group names in Unix and Linux are mapped into numeric forms (UIDs and GIDs, respectively). All file ownership and processes use these numerical names for access control and identity determination throughout the operating system kernel and drivers.

Under many Unix and Linux implementations (via a shadow package), the command `pwck` will perform some simple syntax checking on the password file and can identify some security problems with it. `pwck` will report invalid usernames, UIDs and GIDs, null or nonexistent home directories, invalid shells, and entries with the wrong number of fields (often indicating extra or missing colons and other typos).

Microsoft-Specific Password Issues

Windows uses two password functions—a stronger one designed for Windows NT, 2000, and XP systems, and a weaker one, the LAN Manager hash, designed for backward compatibility with older Windows 9X networking login protocols. The latter is case-insensitive and does not allow passwords to be much stronger than seven characters, even though they may be much longer. These passwords are extremely vulnerable to cracking. On a standard desktop PC, for example, L0phtCrack can try

every short alphanumeric password in a few minutes and every possible keyboard password (except for special ALT-characters) within a few days. Some security administrators have dealt with this problem by requiring stronger and stronger passwords; however, this comes at a cost (see An Argument for Simplified Passwords).

In addition to implementing policies that require users to choose strong passwords, the CERT Coordination Center provides guidelines for securing passwords on Windows systems (CERT, 2002):

Using SYSKEY enables the private password data stored in the registry to be encrypted using a 128-bit cryptographic key. This is a unique key for each system.

By default, the administrator account is never locked out; so it is generally a target for brute force logon attempts of intruders. It is possible to rename the account in User Manager, but it may be desirable to lock out the administrator account after a set number of failed attempts over the network. The NT Resource Kit provides an application called `passprop.exe` that enables Administrator account lockout except for interactive logons on a domain controller.

Another alternative that avoids all accounts belonging to the Administrator group being locked over the network is to create a local account that belongs to the Administrator group, but is not allowed to log on over the network. This account may then be used at the console to unlock the other accounts.

The Guest account should be disabled. If this account is enabled, anonymous connections can be made to NT computers.

The Emergency Repair Disk should be secured, as it contains a copy of the entire SAM database. If a malicious user has access to the disk, he or she may be able to launch a crack attack against it.

Password-Cracking Times

Let us start with a typical password of six characters. When this password is entered into a system's authentication mechanism, the system hashes it and stores the hashed value. The hash, a fixed-sized string derived from some arbitrarily long string of text, is generated by a formula in such a way that it is extremely unlikely that other texts will produce the same hash value—unlikely, but not impossible. Because passwords are not arbitrarily long—they are generally 4 to 12 characters—this reduces the search space for finding a matching hash. In other words, an attacker's password-cracking program does not need to calculate every possible combination of six-character passwords. It only needs to find a hash of a six-character ASCII-printable password that matches the hash stored in the password file or sniffed off the network.

Because an attacker cannot try to guess passwords at a high rate through the standard user interface (as mentioned earlier, the time to enter them is prohibitive, and most systems can be configured to lock the user out after too many wrong attempts), one may assume that the attacker will get them either by capturing the system password file or by sniffing (monitoring communications) on a network segment. Each character in a password is a byte. One does not typically need to consider characters with a leading zero in the highest-order bit, because printable ASCII characters are in codes 32 through 126. ASCII codes 0–31 and 127 are unprintable characters, and 128–255 are special ALT-characters that are not generally used for passwords. This leaves 95 printable ASCII characters.

If there are 95 possible choices for each of the six password characters, this makes the password space $95^6 = 735,091,890,625$ combinations. Modern computers are capable of making more than 10 billion calculations per second. It has been conjectured that agencies such as the NSA have password-cracking machines (or several machines working in parallel) that could hash and check passwords at a rate of 1 billion per second. How fast could an attacker check every possible combination of six-character passwords? 735,091,890,625/1,000,000,000 = about 12 minutes (see Table 2).

What if the system forces everyone to use a seven-character password? Then it would take the attacker 19 hours to brute-force every possible password. Many Windows networks fall under this category. Due to the LAN Manager issue, passwords on these systems cannot be much stronger than seven characters. Thus, it can be assumed that any password sent on a Windows system using LAN Manager can be cracked within a day. What if the system enforces eight-character passwords? Then it would take 77 days to brute-force them all. If a system's standard policy is to require users to change passwords every 90 days, this may not be sufficient.

PASSWORD LENGTH AND HUMAN MEMORY

Choosing a longer password does not help much on systems with limitations such as the LAN Manager hash issue. It also does not help if a password is susceptible to a dictionary or hybrid attack. It only works if the password appears to be a random string of symbols, but that can be difficult to remember. A classic study by psychologist George Miller showed that humans work best with the magic number 7 (plus or minus 2). So it stands to reason that once a password exceeds nine characters, the user is going to have a hard time remembering it (Miller, 1956).

Here is one idea for remembering a longer password. Security professionals generally advise people never to write down their passwords. But the user could write down half of it—the part that looks like random letters and numbers—and keep it in a wallet or desk drawer. The other part could be memorized—perhaps it could be a misspelled dictionary word or the initials for an acquaintance, or something similarly memorable. When concatenated together, the resulting password could be much longer than nine characters, and therefore presumably stronger.

Some researchers have asserted that the brain remembers images more easily than letters or numbers. Thus, some new schemes use sequences of graphical symbols for passwords. For example, a system called PassFace, developed by RealUser, replaces the letters and numbers in passwords with sequences or groups of human faces. It

Table 2 Password Cracking Times

Number of Chars in Password	Number of Possible Combinations of 95 Printable ASCII Chars	Time to Crack (in hours)[a]	Number of Possible Combinations of All 256 ASCII Chars	Time to Crack (in hours)[a]
0	1	0.0	1	0.0
1	95	0.0	256	0.0
2	9025	0.0	65536	0.0
3	857375	0.0	16777216	0.0
4	81450625	0.0	4294967296	0.0
5	7737809375	0.0	1099511627776	0.3
6	735091890625	0.2	281474976710656	78.2
7	69833729609375	19.4	72057594037927900	20016.0
8	6634204312890620	1842.8	18446744073709600000	5124095.6
9	6.E+17	2.E+05	5.E+21	1.E+09
10	6.E+19	2.E+07	1.E+24	3.E+11
11	6.E+21	2.E+09	3.E+26	9.E+13
12	5.E+23	2.E+11	8.E+28	2.E+16
13	5.E+25	1.E+13	2.E+31	6.E+18
14	5.E+27	1.E+15	5.E+33	1.E+21
15	5.E+29	1.E+17	1.E+36	4.E+23
16	4.E+31	1.E+19	3.E+38	9.E+25

[a]Assume 1 billion hash & check operations/second.

is one of several applications that rely on graphical images for the purpose of authentication. Another company, Passlogix, has a system in which users can mix drinks in a virtual saloon or concoct chemical compounds using an onscreen periodic table of elements as a way to log onto computer networks.

AN ARGUMENT FOR SIMPLIFIED PASSWORDS

Employing all of the guidelines for a strong password (length, mix of upper and lower case, numbers, punctuation, no dictionary words, no personal information, etc.) as outlined in this chapter may not be necessary after all.

This is because, according to security expert and TruSecure Chief Technology Officer Peter Tippett, statistics show that strong password policies only work for smaller organizations (Tippett, 2001). Suppose a 1,000-user organization has implemented such a strong password policy. On average, only half of the users will actually use passwords that satisfy the policy. Perhaps if the organization frequently reminds its users of the policy, and implements special software that will not allow users to have "weak" passwords, this figure can be raised to 90%. It is rare that such software can be deployed on all devices that use passwords for authentication; thus there are always some loopholes. Even with 90% compliance, this still leaves 100 easily guessed User/ID password pairs. Is 100 better than 500? No, because either way, an attacker can gain access. When it comes to strong passwords, anything less than 100% compliance allows an attacker entré to the system.

Second, with modern processing power, even strong passwords are no match for current password crackers. The combination of 2.5-gigahertz clock speed desktop computers and constantly improving hash dictionaries and algorithms means that, even if 100% of the 1,000 users had passwords that met the policy, a password cracker might still be able to defeat them. Although some user ID/password pairs may take days or weeks to crack, approximately 150 of the 1000, or 15%, can usually be bruteforced in a few hours.

In addition, strong passwords are expensive to maintain. Organizations spend a great deal of money supporting strong passwords. One of the highest costs of maintaining IT help desks is related to resetting forgotten user passwords. Typically, the stronger the password (i.e., the more random), the harder it is to remember. The harder it is to remember, the more help desk calls result. Help desk calls require staffing, and staffing costs money. According to estimates from such technology analysts as the Gartner Group and MetaGroup, the cost to businesses for resetting passwords is between $50 and $300 per computer user each year (Salkever, 2001).

So, for most organizations, the following might be a better idea than implementing strong password policy: Simply recognize that 95% of users could use simple (but not basic) passwords—that is, good enough to keep a casual attacker (not a sophisticated password cracker) from guessing them within five attempts while sitting at a keyboard. This could be four or five characters (no names or initials), and changed perhaps once a year. In practical terms, this type of password is equivalent to the current "strong" passwords. The benefit is that it is much easier and cheaper to maintain.

Under this scenario, a system could still reserve stronger passwords for the 5% of system administrators who wield extensive control over many accounts or devices. In addition, a system should make the password file very difficult to steal. Security administrators should

also introduce measures to mitigate sniffing, such as network segmentation and desktop automated inventory for sniffers and other tools. Finally, for strongest security, a system could encrypt all network traffic with IPSec on every desktop and server.

Dr. Tippett states: "If the Promised Land is robust authentication, you can't get there with passwords alone, no matter how 'strong' they are. If you want to cut costs and solve problems, think clearly about the vulnerability, threat and cost of each risk, as well as the costs of the purported mitigation. Then find a way to make mitigation cheaper with more of a security impact" (Tippett, 2001).

CONCLUSION

Passwords have been widely used in computing systems since the 1960s; password security issues have followed closely behind. Now, the increased and very real threat of cybercrime necessitates higher security for many networks that previously seemed safe. Guaranteeing accountability on networks—i.e., uniquely identifying and authenticating users' identities—is a fundamental need for modern e-commerce. Strengthening password security should be major goal in an organization's overall security framework. Basic precautions (policies, procedures, filtering mechanisms, encryption) can help reduce risks from password weaknesses. However, lack of user buy-in and the rapid growth of sophisticated cracking tools may make any measure taken short-lived. Additional measures, such as biometrics, certificates, tokens, smart cards, and other means can be very effective for strengthening authentication, but the tradeoff is additional financial burden and overhead. It is not always an easy task to convince management of inherent return on these technologies, relative to other system priorities. In these instances, organizations must secure their passwords accordingly and do the best they can with available resources.

GLOSSARY

Access control The process of limiting access to system information or resources to authorized users.

Accountability The property of systems security that enables activities on a system to be traced to individuals who can then be held responsible for their actions.

ARPANET The network first constructed by the Advanced Research Projects Agency of the U.S. Department of Defense (ARPA), which eventually developed into the Internet.

Biometrics Technologies for measuring and analyzing living human characteristics, such as fingerprints, especially for authentication purposes. Biometrics are seen as a replacement for or augmentation of password security.

Birthday paradox The concept that it is easier to find two unspecified values that match than it is to find a match to some particular value. For example, in a room of 25 people, if one person tried to find another person with the same birthday, there would be little chance of a match. However, there is a very good chance that some pair of people in the room will have the same birthday.

Brute force A method of breaking decryption by trying every possible key. The feasibility of a brute-force attack depends on the key length of the cipher and on the amount of computational power available to the attacker. In password cracking, tools typically use brute force to crack password hashes after attempting dictionary and hybrid attacks to try every remaining possible combination of characters.

CERT Computer Emergency Response Team. An organization that provides Internet security expertise to the public. CERT is located at the Software Engineering Institute, a federally funded research and development center operated by Carnegie Mellon University. Its work includes handling computer security incidents and vulnerabilities and publishing security alerts.

Cipher A cryptographic algorithm that encodes units of plaintext into encrypted text (or *ciphertext*) through various methods of diffusion and substitution.

Ciphertext An encrypted file or message. After plaintext has undergone encryption to disguise its contents, it becomes ciphertext.

Crack, cracking Traditionally, using illicit (unauthorized) actions to break into a computer system for malicious purposes. More recently, either the art or science of trying to guess passwords, or copying commercial software illegally by breaking its copy protection.

CTSS Compatible Time Sharing System. An IBM 7094 timesharing operating system created at MIT Project MAC and first demonstrated in 1961. May have been the first system to use passwords.

Dictionary attack A password cracking technique in which the cracker creates or obtains a list of words, names, etc., derives hashes from the words in the list, and compares the hashes with those captured from a system user database or by sniffing.

Entropy In information theory, a measure of uncertainty or randomness. The work of Claude Shannon defines it in bits per symbol.

Green Book The 1985 U.S. DoD CSC-STD-002–85 publication *Password Management Guideline*, which defines good practices for safe handling of passwords in a computer system.

Hybrid attack A password-cracking technique that usually takes place after a dictionary attack. In this attack, a tool will typically iterate through its word list again using adding certain combinations of a few characters to the beginning and end of each word prior to hashing. This attempt gleans any passwords that a user has created by simply appending random characters to a common word.

Kerberos A network authentication protocol developed at MIT to provide strong authentication for client/server applications using secret-key cryptography. It keeps passwords from being sent in the clear during network communications and requires users to obtain "tickets" to use network services.

MAC Message authentication code, a small block of data derived by using a cryptographic algorithm and secret key that provide a cryptographic checksum for the input data. MACs based on cryptographic hash functions are known as HMACs.

Moore's Law An observation named for Intel cofounder Gordon Moore that the number of transistors per square inch of an integrated circuit has doubled every year since integrated circuits were invented. This "law" has also variously been applied to processor speed, memory size, etc.

Nonce A random number that is used once in a challenge–response handshake and then discarded. The one-time use ensures that an attacker cannot inject messages from a previous exchange and appear to be a legitimate user (see Replay Attack).

One-way hash A fixed-sized string derived from some arbitrarily long string of text, generated by a formula in such a way that it is extremely unlikely that other texts will produce the same hash value.

One-time password Also called OTP. A system that requires authentication that is secure against passive attacks based on replaying captured reusable passwords. In the modern sense, OTP evolved from Bellcore's S/KEY and is described in RFC 1938.

Orange Book 1983 U.S. DoD 5200.28-STD publication, *Trusted Computer System Evaluation Criteria*, which defined the assurance requirements for security protection of computer systems processing classified or other sensitive information. Superseded by the Common Criteria.

Password synchronization A scheme to ensure that a known password is propagated to other target applications. If a user's password changes for one application, it also changes for the other applications that the user is allowed to log onto.

Plaintext A message or file to be encrypted. After it is encrypted, it becomes *ciphertext*.

Promiscuous mode A manner of running a network device (especially a monitoring device or sniffer) in such a way that it is able to intercept and read every network packet, regardless of its destination address. Contrast with nonpromiscuous mode, in which a device only accepts and reads packets that are addressed to it.

Replay attack An attack in which a valid data transmission is captured and retransmitted in an attempt to circumvent an authentication protocol.

Salt A random string that is concatenated with a password before it is operated on by a one-way hashing function. It can prevent collisions by uniquely identifying a user's password, even if another user has the same password. It also makes hash-matching attack strategies more difficult because it prevents an attacker from testing known dictionary words across an entire system.

SAM Security Account Manager. On Windows systems, the secure portion of the system registry that stores user account information, including a hash of the user account password. The SAM is restricted via access control measures to administrators only and may be further protected using SYSKEY.

Shadow password file In the Unix or Linux, a system file in which encrypted user passwords are stored so they are inaccessible to unauthorized users.

Single sign-on A mechanism whereby a single action of user authentication and authorization can permit a user to access all computers and systems on which that user has access permission, without the need to enter multiple passwords.

Sniffing The processes of monitoring communications on a network segment via a wire-tap device (either software or hardware). Typically, a sniffer also has some sort of "protocol analyzer" which allows it to decode the computer traffic on which it's eavesdropping and make sense of it.

Social engineering An outside hacker's use of psychological tricks on legitimate users of a computer system, in order to gain the information (e.g., user IDs and passwords) needed to gain access to a system.

SSH Secure Shell. An application that allows users to login to another computer over a network and execute remote commands (as in rlogin and rsh) and move files (as in ftp). It provides strong authentication and secure communications over unsecured channels.

SSL Secure Sockets Layer. A network session layer protocol developed by Netscape Communications Corp. to provide security and privacy over the Internet. It supports server and client authentication, primarily for HTTP communications. SSL is able to negotiate encryption keys as well as authenticate the server to the client before data is exchanged.

SYSKEY On Windows systems, a tool that provides encryption of account password hash information to prevent administrators from intentionally or unintentionally accessing these hashes using system registry programming interfaces.

CROSS REFERENCES

See *Authentication; Biometric Authentication; Computer Security Incident Response Teams (CSIRTs); Digital Signatures and Electronic Signatures; Disaster Recovery Planning; Encryption; Guidelines for a Comprehensive Security System; Public Key Infrastructure (PKI); Secure Sockets Layer (SSL).*

REFERENCES

Brotzman, R. L. (1985). *Password management guideline* (Green Book). Fort George G. Meade, MD: Department of Defense Computer Security Center.

Brown, A. (2002). *U.K. study: Passwords often easy to crack*. Retrieved 2002 from CNN.com Web site: http://www.cnn.com/2002/TECH/ptech/03/13/dangerous.passwords/index.html

CERT Coordination Center (2002). *Windows NT configuration guidelines*. Retrieved 2002 from CERT Web site: http://www.cert.org/tech_tips/win_configuration_guidelines.html

Federation of American Scientists (FAS) (n.d.). Retrieved May 16, 2003, from www.fas.org/irp/congress/2000_hr/030200_mitnick.htm

Latham, D. C. (1985). *Trusted computer system evoluation criteria* (Orange Book). Fort George G. Meade, MD: Department of Defense National Computer Security Center.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *The Psychological Review, 63*, 81–97.

Salkever, A. (2001). *Picture this: A password you never*

forget. Retrieved 2001 from BusinessWeek.com Web site: http://www.businessweek.com/bwdaily/dnflash/may2001/nf20010515_060.htm

Sloane, N. J. A., & Wyner, A. D. (Eds.). (1993). *Claude Elwood Shannon*: *Collected papers*. New York: IEEE Press.

Tippett, P. (2001). Stronger passwords aren't. *Information Security*. Retrieved 2001 from TruSecure Corporation Web site: http://www.infosecuritymag.com/articles/june01/columns_executive_view.shtml

US v. ElcomSoft & Sklyarov FAQ. Retrieved May 23, 2003, from http://www.eff.org/IP/DMCA/US_v_Elcomsoft/us_v_elcomsoft_faq.html

Zatko, P. "Mudge" (1999b). *Vulnerabilities in the S/KEY one time password system*. Retrieved 1999 from L0pht Heavy Industries, Inc., Web site: http://www.unix.geek.org.uk/~arny/junk/skeyflaws.html

FURTHER READING

Barbalace, R. J. (1999), *How to choose a good password (and why you should)*. Retrieved 1999 from MIT Student Information Processing Board Web site: http://www.mit.edu/afs/sipb/project/doc/passwords/passwords.html

Bobby, P. (2000). *Password cracking using focused dictionaries*. Retrieved 2000 from the SANS Institute Web site: http://rr.sans.org/authentic/cracking.php

Botzum, K. (2001). *Single sign on—A contrarian view*. Retrieved 2001 from IBM Software Services for WebSphere Web site: http://www7b.software.ibm.com/wsdd/library/techarticles/0108_botzum/botzum.html

Cain & Abel [computer software]. Retrieved from http://www.oxid.it

Cracklib and associated PAM modules [computer software]. Retrieved from http://www.kernel.org/pub/linux/libs/pam/Linux-PAM-html/pam.html

Curry, D. A. (1990). *Improving the security of your Unix system*. Retrieved 1990 from Information and Telecommunications Sciences and Technology Division, National Institutes of Health Web site: http://www.alw.nih.gov/Security/Docs/unix-security.html

Cyberspace History List-Server (CYHIST). Retrieved from http://maelstrom.stjohns.edu/archives/cyhist.html

Donovan, C. (2000). *Strong passwords*. Retrieved 2000 from the SANS Institute Web site: http://rr.sans.org/policy/password.php

Elcomsoft [computer software]. Retrieved from http://www.elcomsoft.com/prs.html

Frisch, A. (1999). *Essential system administration* (2nd ed.). Sebestopol, CA: O'Reilly & Associates.

Intertek [computer software]. Retrieved from http://www.intertek.org.uk/downloads

Jablon, D. P. (1997). Extended password key exchange protocols immune to dictionary attack. In *Proceedings of the 6th Workshop on Enabling Technologies* Infrastructure for Collaborative Enterprises, Institute of Electrical and Electronics Engineers, Inc. Retrieved 1997 from http://www.computer.org/proceedings/wetice/7967/79670248abs.htm

John the Ripper [computer software]. Retrieved from http://www.openwall.com/john

LC4 (L0phtcrack 4)[computer software]. Retrieved from http://www.atstake.com

Litchfield, D. (2002). *Hackproofing Oracle application server (A guide to securing Oracle 9)*. Retrieved 2002 from NGSSoftware Web site: http://www.nextgenss.com/papers/hpoas.pdf

Luby, M. and Rackoff, C. (1989). A study of password security. *Journal of Cryptology, 1* (3), 151–158.

McCullagh, D. (2001). *Russian Adobe hacker busted*. Retrieved 2001 from Wired.com Web site: http://www.wired.com/news/politics/0,1283,45298,00.html

McGraw, G., and Viega, J. (2000). *Protecting passwords: Part 1*. Retrieved 2000 from IBM Web site: http://www-106.ibm.com/developerworks/security/library/pass1/index.html?dwzone=security Microsoft Personal Security Advisor [computer software], retrieved from www.microsoft.com/security/mpsa

Morris, R. T., and Thompson, K. (1979). Password security: A case history. *Communications of the ACM, 22*(11), 594–597.

Netscape (2002). *Choosing a good password*. Retrieved 2002 from Netscape Web site: http://www.netscape.com/security/basics/passwords.html

Nomad, S. (1997). *The unofficial NT hack FAQ*. Retrieved 1997 from http://www.nmrc.org/faqs/nt/index.html

Nordahl-Hagen, P. NET Password Recovery [computer software]. Retrieved from http://home.eunet.no/~pnordahl/ntpasswd/

Npasswd [Computer software for SunOS 4/5, Digital Unix, HP/UX, and AIX]. Retrieved from http://www.utexas.edu/cc/unix/software/npasswd

Pandora [computer software]. Retrieved from http://www.nmrc.org/pandora

Passfilt [computer software]. Retrieved from http://support.microsoft.com/support/kb/articles/Q161/9/90.asp

Passlogix [computer software]. Retrieved from http://www.passlogix.com

Raymond, E. S. (1999). A brief history of hackerdom. Retrieved 1999 from Eric. S. Raymond Web site: http://tuxedo.org/~esr/writings/hacker-history/

RealUser [computer software]. Retrieved from http://www.realuser.com

Russell, R. (Ed.) (2002). *Hack proofing your network*. Syngress Publishing.

Sabin, T. PWDUMP2 [computer software]. Retrieved from http://www.webspan.net/~tas/pwdump2

Sanjour, J., Arensburger, A., and Brink, A. (2000). *Choosing a good password*. Retrieved 2000 from Computer Science Department, University of Maryland Web site: http://www.cs.umd.edu/faq/Passwords.shtml

SANS Institute (2002). *The twenty most critical Internet security vulnerabilities*. Retrieved 2002 from The Sans Institute Web site: http://www.sans.org/top20.htm

Schneier, B. (2000). *Secrets & lies: Digital security in a networked world*. New York: Wiley.

Smith, R. E. (2001). Authentication: From passwords to public keys. Boston: Addison Wesley Longman.

The FBI/SANS Twenty Most Critical Internet Security Vulnerabilities. Retrieved from http://www.sans.org/top20.htm

Zatko, P. "Mudge" (1999a). *L0phtCrack 2.5 Readme.doc* L0pht Heavy Industries, Inc. [now @stake, Inc.]

Patent Law

Gerald Bluhm, *Tyco Fire & Security*

INTRODUCTION

This chapter introduces the fundamental concepts of patent law, both in the United States and internationally, with some focus on software and Internet-related issues. Patents have been described as monopolies for limited terms, in exchange for inventors disclosing how their inventions are made or used. With the promise of such monopolies, inventors are encouraged to invent and thus reap the rewards made possible by the rights accorded. Competitors must either obtain a license to make or use a patented invention or discover new ways that circumvent a patented invention as defined by the patent claims.

Some have rejected the use of the word "monopoly" to describe patents. Regardless of whether one uses the word "monopoly," certain rights are granted to the owner of a patent: the right to *exclude* others from making, using, selling, or offering to sell the invention in the United States, importing the invention into the United States, or importing into the United States something made by a patented process. What may not be obvious is that a patent does not grant its owner the right to make, use, sell, offer to sell, or import the patented invention. In fact, many patented inventions are improvements made on existing (and patented) work, and if made, used or sold, they would constitute infringement of the earlier patent.

U.S. PATENT LAW

Constitutional Basis

The U.S. Constitution grants to Congress the power "To promote the Progress of . . . useful Arts, by securing for limited Times to . . . Inventors the exclusive Right to their respective . . . Discoveries" (U.S. Constitution Article I, Section 8, Clause 8). In accordance with this power, Congress has over time enacted several patent statutes. In particular, in 1952, the present patent law, codified under Title 35 of the United States Code (abbreviated as "35 U.S.C.," available on the Web at http://uscode.house.gov/title_35.htm), was enacted, although it has been amended many times over the years.

How Does an Inventor Get a Patent?

The U.S. Patent and Trademark Office

Under the U.S. Department of Commerce, the U.S. Patent and Trademark Office (USPTO; http://www.uspto.gov) processes patent applications and ultimately issues or grants patents. During the processing of an application (a process known as *patent prosecution*), the application is examined by an examiner who is familiar with the specific technology field of the invention described in the application. Typically, the examiner will object to the application because he or she feels that, when compared with prior art (existing knowledge possessed or information accessible by those in the subject technology field), there is nothing novel or unobvious about the invention. Patent prosecution typically involves communications back and forth between the examiner and the inventor (or the inventor's patent attorney or agent) in which the inventor or attorney clarifies for the examiner how the invention is in fact novel and unobvious over the prior art.

Inventors can represent themselves before the USPTO. Alternatively, an inventor (or the assignee to whom the inventor assigns ownership of an invention) may employ an attorney or agent registered with the USPTO. Both patent attorneys and patent agents have technical backgrounds in some science or engineering field and have taken and passed a registration examination administered by the USPTO. In addition, patent attorneys have completed law school and are admitted to practice law in at least one jurisdiction, whereas patent agents are not attorneys.

What Is Patentable?

An inventor may obtain a patent for "any new and useful process, machine, manufacture, or composition of matter, or any new and useful improvement thereof" (35 U.S.C. §101). In a landmark U.S. Supreme Court case in which whether a live, human-made microorganism could be patented was at issue, the Supreme Court unequivocally stated that "anything under the sun that is made by man" is patentable (*Diamond v. Chakrabarty,* 1980).

What is not patentable? Generally speaking, laws of nature, physical phenomena and abstract ideas, per se, are not patentable. For example,

> a new mineral discovered in the earth or a new plant found in the wild is not patentable.... Likewise, Einstein could not patent his celebrated law that $E = mc^2$; nor could Newton have patented the law of gravity. (*Diamond v. Chakrabarty*, 1980)

Of course, a practical application of some physical phenomena may be patentable. For example, although a new plant found in the wild is not patentable, its medicinal use may be patentable. In *Diamond v. Chakrabarty* (1980), the Supreme Court clearly stated that even living things, in this case microorganisms produced by genetic engineering, are patentable. In fact, all that matters is whether the living matter is the result of human intervention.

In particular, Internet-related inventions are patentable and can be protected with method claims, apparatus claims, so-called Beauregard claims, embedded signal claims, and so on, all of which are discussed later in the chapter. Many Internet-related inventions are protected by "business method" patents.

Business Methods

Prior to the *State Street Bank and Trust Co. v. Signature Financial Group, Inc.* (1998) decision by the Court of Appeals for the Federal Circuit (CAFC) in 1998, there was some uncertainty as to whether methods of doing business were patentable. Although the invention claimed was technically a "machine" that implemented business methods, *State Street* (available at Georgetown University's Web site at http://www.ll.georgetown.edu/federal/judicial/fed/opinions/97opinions/97-1327.html) is cited for confirming that indeed business methods themselves are patentable.

In that decision, Signature was the assignee (owner) of U.S. Patent No. 5,193,056. The claimed invention is a system in which mutual funds pool their assets into an investment portfolio to take advantage of economies of scale in administering investments. State Street Bank had been negotiating a license with Signature Financial. When negotiations broke down, State Street Bank sought a declaratory judgment that the patent was invalid because it described a business method. The court, however, determined that indeed business methods are patentable subject matter.

When is a method a business method? It is not always clear if a particular method is strictly a business method. For example, Amazon.com received a patent (U.S. Patent No. 5,960,411) for its 1-click invention. Although the invention has been labeled a method of doing business by some, Amazon has asserted that its 1-click patent is not a business method patent.

The USPTO has established a classification system, with more than 400 classes, which are further divided into subclasses. Every application is assigned to a class and subclass according to the technology of the invention. In general, methods that fall into the USPTO's Class 705 ("Data processing: financial, business practice, management, or cost/price determination") are considered to be business methods. Refer to http://www.uspto.gov/web/offices/ac/ido/oeip/taf/def/705.htm for a list of Class 705 categories. For example, some of the first few subcategories of Class 705 include health care management; insurance; reservation, check-in, or booking display for reserved space; staff scheduling or task assignment; market analysis, demand forecasting, or surveying; and so on. For those inventions that are considered to be business methods, some special rules apply during prosecution of a business method patent application and with respect to infringement. Whereas most patent applications are subjected to examination by a single examiner, the USPTO subjects business method applications to one or more extra reviews. This extra review was added in part as a response to numerous complaints, made in the popular press and elsewhere, that many patents were being issued on inventions that were "clearly" not patentable.

Furthermore, accused infringers of issued business method patents have at their disposal an extra defense against the accusation that infringers of other types of patents do not have. For example, for most patents, if Party A receives a patent for an invention, and Party B has been practicing the invention before issuance of the patent, Party B must stop its practice or obtain a license once the patent issues. After *State Street*, however, Congress added §273 to Title 35 of the U.S. Code as part of the American Inventor's Protection Act of 1999, providing for "intervening rights" to protect parties that may not have applied for a business method patent based on the misconception that such patents were unobtainable. The details of §273 are beyond the scope of this article, but basically it provides, in certain situations, a defense to an infringement claim for a party that was using the patented business method before the patent issued.

Requirements of an Invention

Three basic requirements must be met before one can obtain a patent for an invention: the invention must be novel, it must not be obvious in view of the current state of the art (and with respect to a person knowledgeable or "of ordinary skill" in the art), and it must be useful.

Novelty is statutorily provided for in 35 U.S.C. §102, which describes several conditions in which a patent may not be obtained: 35 U.S.C. §102 states that a patent cannot be obtained if the invention was known or used by others in the United States prior to the patent applicant's invention (this could happen, for example, when two people separately invent the same invention, each unaware of the other's activity or accomplishment) or if the invention has been patented or described in a printed publication anywhere in the world. "Printed publications" may include any information that is freely accessible via the Web, even though a Web page is not technically printed in hard copy.

Even inventors' own writings can be held against them. Inventors have 1 year to file a patent application in the United States if the inventions were patented or described in a printed publication anywhere in the world or if the inventions were in public use or on sale in the U.S.

The prohibition against obviousness is statutorily provided for in 35 U.S.C. §103(a), which states in essence

that even if the invention is not exactly the same as that described in a patent or publication or that is in use or on sale, if the difference is obvious, a patent cannot be obtained. Examiners often provide rejections based on their sense that it would be obvious to combine two or more published patents or other publications that complement each other. Of course, such a combination must be obvious to a person having ordinary skill in the art, and it must have been obvious at the time the invention was made. (Often, by the time a patent issues 2 or 3 years after the patent application was filed, or even when the application was filed, it may seem to be obvious in light of the prior art. However, the critical time to examine obviousness is when the invention was made. As many court decisions show, it is not always an easy thing to cast away current knowledge and place oneself back to the time the invention was made to determine whether it was obvious.) Thus, assertions as to whether an invention is obvious or not in view of the cited art can be highly subjective.

U.S.C. §101 requires that an invention be useful. At least three categories of subject matter have been identified by the Supreme Court that are not, by themselves, patentable: laws of nature, natural phenomena, and abstract ideas. The invention must be useful, concrete, and tangible. For example, the CAFC in *State Street Bank* said that mathematical algorithms by themselves are unpatentable because "they are merely abstract ideas constituting disembodied concepts or truths that are not 'useful.'"

Generally speaking, the requirement that an invention be "useful" is an extremely low bar to patentability. Nonetheless, an invention can fail the usefulness test if an applicant fails to explain adequately why the invention is useful or if an assertion of utility is not credible. For example, the invention considered in *Newman v. Quigg* (1989) was considered to be a perpetual motion machine and thus found to be inoperative (as going against the laws of thermodynamics). It therefore did meet the usefulness standard.

In addition to these requirements, a specification is required in the patent application that includes a written description and at least one claim. The written description must describe the invention and teach enough about it in sufficient detail so as to enable "one skilled in the art" to make or use the invention. The written description must also describe the "best mode" (i.e., the best way to carry out the invention) known to the inventor, although there is no requirement to point out a specific embodiment of the invention as the best mode. (The first paragraph of 35 U.S.C. §112 discusses the requirements of the written description.) The specification must conclude with at least one claim that is the legal statement defining the invention. Courts looks to the claims when determining whether an accused party is infringing a patent. Claims are discussed in more detail later in the chapter. Finally, 35 U.S.C. §113 sets forth the particular requirements for drawings, which must be supplied as necessary to provide an understanding of the invention.

Patent Prosecution

The process of obtaining a patent, from filing of a patent application to responding to office actions from the USPTO, to paying the issue fee, is referred to as *patent prosecution*. The first step in the patent prosecution process, other than invention itself, is often the conducting of a "prior art" search. There is no obligation on the part of an applicant to do a search (although there is an obligation to report known material information to the PTO). Nonetheless, performing a search is often a good idea. If search results show that the invention is not novel, a long and costly (and possibly doomed) prosecution process can be avoided. Even if the invention still appears to be novel after such a search, oftentimes the search helps the person who ultimately drafts the patent application to focus on those parts that are truly novel, by exposing those aspects that are well known or that have been described in printed publications.

A search for U.S. and European patents and published patent applications can be performed, for example, using respectively the USPTO's online search facility and the European Patent Office (EPO) online search facility (see http://ep.espacenet.com). Most patent offices in other countries have their own Web sites that can be searched. A list of these Web sites can be found at the USPTO's and the EPO's Web sites.

In addition, many useful documents and news items may be found on the Web using standard Web searching facilities. Although these are excellent sources, more extensive (and expensive) searches may be conducted using proprietary databases that may contain articles from hundreds or thousands of trade journals, professional publications, newspapers, magazines, and so on.

The next step is preparing or drafting the patent application. As noted earlier, a patent attorney or agent often does this, although inventors can represent themselves before the PTO. The application is then filed with the PTO. Once received by the PTO, an application is assigned an application number and eventually assigned to an art group, consisting of examiners who are familiar with the particular field to which the application/invention pertains.

Eventually the application is assigned to a specific examiner in the art group, who reviews the application in view of both the results of his or her own prior art search and any material information submitted by the applicant. Typically, the examiner objects to one or more aspects of the application, and in an *Office Action*, rejects one or more of the claims based on the prior art. Or the examiner may object to unclear language in the specification or an informality in the submitted drawings. An Office Action is mailed to the applicant (or his attorney), and the applicant must reply within a certain time frame or the application will be considered abandoned.

The applicant can reply to the Office Action in several ways. For example, the applicant can point out the differences between the invention and the prior art cited by the examiner in the Office Action, stressing that the invention is not taught or even suggested by a knowledge of the prior art. In the reply to the Office Action, the applicant can cancel claims, amend claims for clarity, narrow claims to overcome the examiner's rejections, or even broaden claims. New claims may also be added (at least in response to a first Office Action). Corrections to the specification or drawings may also be made, but in any case,

the applicant is never allowed to introduce new matter into the application.

The examiner often makes a subsequent Office Action "final." Certain rules apply when the applicant replies to a final Office Action—for example, new claims cannot normally be added, and only certain amendments of a limited nature are permitted—but a "final" Office Action is not as final as it sounds.

If some claims are allowed in an Office Action, an applicant can, in his or her reply, cancel the rejected claims, permitting a patent to issue with the allowed claims. A new application, called a "continuation" can then be filed with the rejected claims. (Note that this continuation application must be filed while the parent application is pending, that is, before the parent issues as a patent with the allowed claims).

Alternatively, if no claims are allowed, the applicant, in response to a Final Office Action, may file a *Request for Continued Examination*. For the equivalent cost of filing a new application, the applicant is allowed to continue prosecution without the finality of the final office action. In older cases, filed before May 29, 2000, the applicant may, while the first application is pending, file a continuation-type application, called a *Continued Prosecution Application* and allow the first application to become abandoned.

During the course of prosecution, it may be desirable to file a new set of claims while allowing the original application to proceed. For example, the applicant may determine that aspects of the original application not previously claimed may be worth pursuing and may file the same specification with a different set of claims, claiming priority to the first application. This second application is also known as a continuation application. It has its own filing date, but because it claims priority to the first application, it will expire (under the current statute) 20 years from the application date of the first application (or the date of the earliest application in the priority chain). If new matter is added to the specification of a continuation, for example, an improvement or a new configuration, the new application is called a continuation-in-part (CIP). A patent issuing on a CIP application, like other utility patents, expires 20 years from the first application to which the CIP claims priority.

In some cases, an examiner may decide that the claims of an application really describe two or more inventions, each requiring its own prior art search. In this case, the examiner may issue a *restriction requirement* in which the examiner divides the various claims into different groups, each pertaining to a different invention. The applicant is then required to elect one of the groups and to cancel or amend the remaining claims. The canceled claims can be filed (while the original application is still pending) in one or more applications known as "divisional" applications. As with continuations, each divisional application has its own filing date, but each must claim priority to the parent and therefore has a term of 20 years from the filing date of the parent (or the earliest filed application in the priority chain).

Eventually, the applicant hopes, each application (including parent, continuations, CIPs, divisionals) is allowed. For a given allowed patent application, the applicant pays an issue fee, and soon thereafter the patent issues and is then in force. Although the term of a patent is 20 years from the priority date, "maintenance" fees must be paid at specific intervals from the date of issue or the patent expires. Specifically, these intervals are 3 years and 6 months, 7 years and 6 months, and 11 years and 6 months. A 6-month grace period is available for a surcharge.

Appealing an Examiner's Decision

If the applicant is unsatisfied with the examiner's conclusions as to unpatentability, the applicant can appeal to the Board of Patent Appeals and Interferences within the Patent and Trademark Office. Each appeal is heard by at least three members of the board. An applicant who is unhappy with the board's decision may further appeal to the U.S. Court of Appeals for the Federal Circuit (CAFC). The CAFC makes a decision based only on the record from the appeal to the board. Alternatively, an unhappy applicant may file a civil suit against the director of the USPTO in the U.S. District Court for the District of Columbia. Unlike appeals to the CAFC, new evidence may be presented in addition to the record from the appeal to the board.

Publication and Provisional Rights

Applications filed on or after November 29, 2000, are published roughly 18 months from the priority date, unless the applicant specifically requests nonpublication, certifying at the same time that the invention has not been and will not be the subject of an application filed in another country. Early publication can be requested. Applications filed prior to November 29, 2000, but still pending as of that date are not typically published, but publication may be requested.

If a published application eventually issues as a patent, with claims that are "substantially identical" to those published in the application publication, the owner of the patent may be entitled to a reasonable royalty, from the time of the publication date up to the issue date, from someone who makes, uses, offers for sale, or sells the claimed invention in the United States or who imports the claimed invention into the United States and who has actual notice of the published patent application. The rights to these royalties are known as *provisional rights*.

Patent Term—How Long Does a Patent Last?

Patent protection begins on the day a patent issues. Because of a change in law in 1994 to conform to the Uruguay Round of the General Agreement on Tariffs and Trade (GATT), when a patent expires depends on when the application was filed. Prior to the change in law, the term of a U.S. utility patent was 17 years from the issue date. Now, however, any U.S. utility patents issuing from an application filed on or after June 8, 1995, are valid for 20 years from the priority date, that is, the date of the earliest application to which the application claims priority (the earliest filing date in a chain of continuation and divisional applications). Utility patents which were still in force on June 8, 1995, or applications filed prior to that date but still pending, receive the best of both worlds (with regard to patent term): either 17 years from the issue date or 20 years from the priority date, whichever is later.

Various adjustments and extensions may be available under certain conditions. A discussion of these conditions is beyond the scope of this chapter, however. The patent terms discussed here pertain to U.S. utility and plant patents. U.S. design patents expire after 14 years from the issue date.

How to Read a Patent

A patent is organized into several sections. These include a cover sheet, drawings, a specification and claims. The specification includes a background, a summary, a brief description of the drawings, and a detailed description. The cover sheet includes bibliographical information, a short abstract that briefly describes the invention, and usually a copy of one of the drawings considered to be representative of the invention. Drawings must be provided where necessary for understanding the invention. The background section describes prior art, or the state of the art prior to the patented invention. The summary provides a short synopsis of the invention and often is a regurgitation of the claims in plainer language than the claims. A brief description of the drawings typically follows. Next comes a written description (often labeled as a detailed description) of one or more embodiments of the invention. As previously mentioned, the written description must enable any person skilled in the art to make and use the invention. The description must also set forth the "best mode" contemplated by the inventor, although this best mode need not be pointed out as such. Finally, a set of claims is provided which point out and distinctly claim the protected subject matter. Each claim is written as a single sentence and typically consists of a preamble, a transitional phrase, and a set of limitations. For example, Claim 1 of U.S. Patent No. 6,004,596 ("Sealed crustless sandwich") appears as follows:

> I claim:
>
> 1. A sealed crustless sandwich, comprising:
>
> a first bread layer having a first perimeter surface coplanar to a contact surface;
>
> at least one filling of an edible food juxtaposed to said contact surface;
>
> a second bread layer juxtaposed to said at least one filling opposite of said first bread layer, wherein said second bread layer includes a second perimeter surface similar to said first perimeter surface;
>
> a crimped edge directly between said first perimeter surface and said second perimeter surface for sealing said at least one filling between said first bread layer and said second bread layer;
>
> wherein a crust portion of said first bread layer and said second bread layer has been removed.

The *preamble* is the phrase: "A sealed crustless sandwich." The *transitional phrase* is "comprising." These are followed by five limitations: "a first bread layer," "at least one filling," "a second bread layer," "a crimped edge," and the condition "wherein a crust portion . . . has been removed." For this claim to be infringed, an unauthorized party must make, use, sell, or offer to sell or import into the United States a sandwich product that satisfies every one of these limitations. It is irrelevant that a crustless sandwich may have other components not described in the claim, for example, a cherry on top. As long as some food product meets every one of the limitations listed in Claim 1, that product is said to infringe Claim 1. On the other hand, if a sandwich is lacking some element, such as the crimped edge, it cannot literally infringe (but see the subsequent discussion regarding the doctrine of equivalents).

A first claim is typically written broadly to cover a wide range of variations. Narrower claims often follow that include the limitations of the broad claim, plus additional limitations that limit the scope of the invention recited by these narrower claims. Narrower claims are often written as dependent claims.

For example, Claim 1 above is an independent claim. Claim 2 in the same patent reads:

> 2. The sealed crustless sandwich of claim 1, wherein said crimped edge includes a plurality of spaced apart depressions for increasing a bond of said crimped edge.

Claim 2 is called a "dependent" claim because it *depends from* Claim 1, that is, it includes all of the five limitations of Claim 1, plus the further limitation that the crimped edge include "spaced apart depressions." For a sandwich to infringe this claim, it must meet all of the limitations of Claim 1 *and* Claim 2. One reason for providing additional narrower claims is that often, during litigation of a patent suit, some claims may be found to be invalid. Even though a claim may be invalidated in a court of law (for example, if a publication is presented that predates the patent's priority date and that teaches or suggests one or more of the claims), a narrower claim with additional limitations may still be valid, even if it depends from the invalidated claim.

Another reason for providing additional narrower claims is the so-called doctrine of claim differentiation, under which "two claims of a patent are presumptively of different scope" (*Kraft Foods, Inc. v. Int'l Trading Co.*, 2000). According to this doctrine, if a dependent claim includes a narrowing definition of some limitation of a base claim, then the base claim is presumed to encompass not only the narrow definition, but other embodiments as well. For example, Claim 2 may help to support the proposition that Claim 1 covers crustless sandwiches that do not have depressions that are spaced apart as well as crustless sandwiches that have other kinds of bonding mechanisms.

First-time readers of claims are often puzzled by the seemingly bizarre language and grammar used in claims. Sometimes this language results from the statutory requirement that claims particularly point out and distinctly claim the subject matter that the inventor or applicant regards as his or her invention. Thus, use of a definite article

such as "the" is typically not allowed unless it refers to something already defined in the claim (i.e., there is an "antecedent basis" for the thing to which the claim refers). For this reason, one often sees "a number of (things)" where in normal usage, one would say "the number of (things)."

In addition, use of the word "or" is generally frowned upon because it leaves options open and is therefore not considered to distinctly claim an invention. Thus, one often sees in claims language such as "at least one of [Choice A], [Choice B], *and* [Choice C]," or "any of [Choice A], [Choice B], *and* [Choice C]" where in normal speech, one might say "[Choice A], [Choice B], *or* [Choice C]." Similarly, instead of reciting "one or more of," Claims will more often recite "a plurality of" or "at least one of," leading to even more confusing language later in the claims, such as "the at least one of." Although such language may at first be confusing, an understanding of why these terms are used may help in reading and interpreting a claim.

Another aspect of claiming that can be confusing to the layperson is that often almost the exact same language is recited in two different claims. For example, a patent typically will have a method claim and an apparatus (or system) claim that use parallel language. Remember, however, that the right to exclude may be different between a method and an apparatus or composition of matter.

A limitation in an apparatus claim may also be expressed as a means or step for performing a specified function without the recital of any specific structure. While such a limitation is not always triggered by "means for" (also called "means-plus-function") language, and may even be triggered in the absence of such language, such claims are often added to a patent. A "means for" limitation is construed to cover the corresponding structure, material, or acts described in the specification and their equivalents (see 35 U.S.C. §112, sixth paragraph).

In addition to the more or less standard apparatus, method, and "means-for" claims, computer- and software-related inventions are often additionally recited in so-called Beauregard claims and signal claims. As the USPTO states, a computer program is merely a set of instructions, capable of being executed, but is not itself a process [see section 2106(a) of the Manual of Patent Examining Procedure]. To be patentable, a computer-readable medium, or an "article of manufacture" comprising a computer-readable or useable medium, is claimed, having therein a computer program that performs some steps of a process. These types of claims have been called Beauregard claims after the inventor of one of the first patent applications to use such claims (see Beauregard, n.d.).

Another type of claim one might encounter is the so-called propagated signal claim. Such a claim might appear as follows:

> 1. A computer data signal embodied in a carrier wave for [doing something], the computer data signal comprising:
>
> program code for [performing a first action];
>
> program code for [performing a second action];
>
> etc.

Such claims are thought to protect against the unlicensed transmission of a computer program over a network such as the Internet or through modems. Of special concern with claims directed to client–server applications is the fact that a single party may not be performing or using all of the limitations of a claim. In other words, if a claim recites actions taken by both the server and the client and two independent parties control the server and client, then neither party can be an infringer. Therefore, it may be desirable in a patent to have one set of claims directed to the overall invention, another set of claims directed to actions taken at the server (possibly in response to messages received from a client), and yet another set of claims directed to actions taken at a client (possibly in response to messages received from a server).

Protecting Patent Rights

What Rights Are Conferred on a Patentee?

A patent confers specific "exclusive" rights on the owner of a patent. That is, the owner of a patent is granted the right to exclude other parties from various acts, including making, using, offering to sell, or selling the patented invention (as set forth in the claims) within the United States or importing the patented invention into the United States. Where a process is patented (as opposed to an apparatus or composition of matter), the patentee is similarly granted the right to exclude others from using, offering to sell, or selling in the United States or importing into the United States any product made by the claimed process. Note that a patent does not give an owner the right to practice the invention recited in the patent; another (broader) patent may exclude the owner from practicing the invention. These rights begin when the patent is granted, that is, when the patent issues and last until the patent expires. See the Patent Term section for an explanation of how the term of a patent is calculated.

What Is Infringement?

When someone performs one of the restricted acts regarding patented material or a patented method without permission of the patent owner, the patent is said to be "infringed." There are three types of infringement: direct infringement, active inducement to infringe, and contributory infringement. Direct infringement occurs when someone literally performs one of the restricted acts, that is, makes, uses, sells the invention (as set forth in at least one claim of the patent), or offers it for sale in the United States or imports it into the United States. Note that this does not require knowledge by the infringer that the invention is prohibited, nor does it require that the infringer intentionally perform the act. All that is required for direct infringement is the act.

Active inducement to infringe occurs when somebody induces another to infringe. If there is no direct infringement, there cannot be active inducement to infringe, no matter how hard someone tries to induce infringement. (Of course, there could be other legal issues in this case.) Contributory infringement occurs when a component of a patented invention is sold or offered for sale in the United States, or imported into the United States, by a party who is aware that the component is especially made

or adapted for an infringing use. As with active inducement to infringe, there cannot be any contributory infringement unless there is direct infringement by some party. Note that because use must occur within the United States, in an Internet-related patent, there may be no infringement when either a server or a client outside the United States performs some of the elements of a patent claim.

Doctrine of Equivalents and *Festo*

To infringe a patent, a device must have all of the pieces recited in at least one claim. If the device matches the claim's limitations exactly, it is said to "literally" infringe the patent claim. For method patents, if a process matches all of the steps recited by at least one claim, the method is literally infringed. Even if the accused device or process does not match the claims exactly, the judicially created doctrine of equivalents provides that a device that does not literally infringe may still infringe if the differences are "insubstantial." Arguments made by an applicant to overcome an examiner's rejection during prosecution of a patent application may limit the breadth of equivalents. This principle is known as *prosecution estoppel*. In the past, application of prosecution estoppel was somewhat flexible, depending on many variables. More recently, in a 1999 decision, the CAFC, in *Festo Corporation v. Shoketsu Kinzoku Kogyo Kabushiki Co., Ltd.*, drastically limited the scope of the doctrine of equivalents, such that any change to a claim limitation made for patentability acts as a complete bar to the application of the doctrine, as to that limitation.

Even more recently, however, in May 2002, the U.S. Supreme Court overturned the CAFC's complete bar. Instead, the Supreme Court has held that when a claim is narrowed, equivalence may be barred, but the patentee can overcome this bar by showing that "at the time of the amendment one skilled in the art could not reasonably be expected to have drafted a claim that would have literally encompassed the alleged equivalent" (*Festo*, 535 U.S. (2002), slip opinion available at http://www.supremecourtus.gov/opinions/01pdf/00–1543.pdf).

What Are the Remedies?

When the owner of a patent believes another party is infringing, the owner typically seeks one of two things: to have the accused party cease from engaging in the infringing acts or to license the invention to the accused party in order to collect royalties. An owner typically files a lawsuit seeking one or more remedies if the owner does not wish to license the invention or if license negotiations break down or to "persuade" the accused party to obtain a license. The patent owner can seek an injunction against the accused party in which the court can order a cessation of the infringing act. A court can also award monetary damages to compensate the owner for the infringement. These damages can include a reasonable royalty for the use of the invention, as well as interest and other costs on which the court may decide. A court may increase these monetary damages up to three times. In exceptional cases, a court may award attorney fees. (This can be for either party, whichever prevails.)

Defending Against an Accusation of Infringement

Several defenses are available to a party being accused of infringement. This is beyond the scope of this chapter, but a list of possible defenses includes noninfringement, absence of liability, unenforceability, and invalidity of the patent based on numerous reasons. Issuance of a patent by the USPTO creates a presumption of validity of the patent. Nonetheless, a court may rule, based on the patent itself, the prosecution history of the patent application, or new evidence, that the patent is not valid. Generally speaking, damages cannot be obtained for infringements that occurred earlier than 6 years before the filing of the suit. Independent of the statutory time frame, an infringement suit may be barred by *laches,* for example, if the patent owner deliberately delayed bringing the suit for an unreasonable time, knowing that the delay will work to the detriment of the accused infringer. Recently, the CAFC confirmed the existence of *prosecution laches,* in which an unreasonable delay during prosecution of the application, together with harm to the other party caused by the delay, can result in unenforceability of a patent *Symbol Technologies, Inc. et al. v. Lemelson Medical, Education & Research Foundation, Limited Partnership,* 2002).

Court Jurisdiction in Patent Cases

Because patent law is federal law (as opposed to state law), federal courts have jurisdiction over all patent-related cases. Furthermore, although there are many federal courts of appeal, Congress, in seeking to establish a single interpretation of the patent laws, established the Court of Appeals for the Federal Circuit to hear all patent-related appeals (but see below, re *Holmes Group, Inc.*).

Typically, a three-judge panel hears and decides an appealed case. On occasion, that decision may be appealed. One of the parties may ask for an *en banc* rehearing in which all or most of the judges from the CAFC rehear the case. If a party is still not satisfied with the final ruling, they may appeal to the U.S. Supreme Court. Few patent cases are ever heard by the Supreme Court, however, typically one or two a year, if that.

One case that was heard by the Supreme Court recently was *Holmes Group, Inc. v. Vornado Air Circulation Systems, Inc.* (2002), (slip opinion at http://www.supremecourtus.gov/opinions/01pdf/01–408.pdf). In this case, Holmes filed a complaint seeking a declaratory judgment that their products did not infringe Vornado's *trade dress*. Trade dress is another form of protection that does not involve patents. Although the original complaint did not involve patents, Vornado filed a counterclaim alleging patent infringement.

On appeal (after an appeal to and decision by the CAFC), the Supreme Court ruled that the CAFC does not have jurisdiction over a case that involves questions of patent law in which the party bringing the suit did not, in its complaint, assert any patent law issues. Thus, although the CAFC was created in part to form a uniform interpretation of patent law across the country, where the original complaint does not assert any patent law issues, the CAFC does not have jurisdiction, even where patent issues are later asserted in a counterclaim.

Reasons for Obtaining a Patent

There are many reasons a party may wish to obtain patent protection, and the brief discussion presented here is not intended to be all-inclusive. One reason for obtaining a patent is to protect one's intellectual property. Thus, the holder of the patent may be able to prevent competition by preventing others from taking advantage of the invention. Others may decide not to compete at all in the particular area, or may decide to spend significant amounts of time and money developing processes or products that do not infringe. Alternatively, if both sides are willing, the patent holder may license part or all of the patent to another party for a fixed fee, a royalty, or some combination of the two. A license is basically an agreement between a licensor (patent holder) and licensee that the licensor will not sue the licensee for what would otherwise constitute infringement of some or all of the patent claims.

Another reason for obtaining a patent is more defensive. For example, Company B may be reluctant to sue Company A for infringement of Company B's patent, if Company B thinks that Company A may countersue for infringement of Company A's patents. Often, such a situation may result in cross-licensing between the two parties, in which each agrees not to sue the other for infringement of all or part of each other's patents.

Yet another reason for obtaining a patent, especially for start-up companies, is for attracting investment. Investors like to know that there is some value in whatever they are investing in, and the existence of one or more patents (or even pending patent applications) may be an indication of a company's viability.

Types of Patents

In the United States, there are several types of patents: utility patents, plant patents, and design patents. Utility patents are the patents that most people typically think of when they think of patents. Utility patents may be obtained for "any new and *useful* process, machine, manufacture, or composition of matter, or any new and useful improvement thereof" (35 U.S.C. §101). The Plant Patent Act of 1930, now codified as 35 U.S.C. §161, allows "plant patents" for plants that are asexually reproduced. Tubers, such as potatoes, are excluded (i.e., they are not patentable). To be patentable, a plant must have been found in an uncultivated state (e.g., *not* in a garden). A plant patent includes the right to exclude others from asexually reproducing the plant; using, offering for sale, or selling the plant (or parts of the plant) in the United States; or importing the plant into the United States if it was asexually reproduced (see 35 U.S.C. §§161–164).

Despite the availability of plant patents, utility patents may also be obtained for both sexually and asexually reproduced plants (see *J.E.M. AG Supply Inc, dba Farm Advantage, Inc., et al. v. Pioneer Hi-Bred International, Inc.*, 2001; slip op at www.supremecourtus.gov/opinions/01pdf/99–1996.pdf). The requirements for obtaining a plant patent are more relaxed than the requirements for obtaining a utility patent, however.

Design patents may be obtained for new, original, and ornamental designs for manufactured articles. A design patent protects the way an article looks, as depicted in the drawings. Design patents have a term of 14 years from the issue date. Design patents may be obtained for computer-generated icons, including full-screen displays and individual icons. The USPTO's Manual of Patent Examining Procedure (MPEP), section 1504.01(a) provides "Guidelines for Examination of Design Patent Applications for Computer-generated Icons." (See U.S. Design Patent D453,769 for an example of a design patent for a computer-generated icon.) According to the MPEP (section 1504.01(a).1.A), to satisfy the manufactured article requirement, such an icon must be claimed as "a computer-generated icon shown on a computer screen, monitor, other display panel, or a portion thereof" or with similar language. The icon must also be fundamentally ornamental rather than functional. Fonts may also be patented with design patents. (For an example, see U.S. Design Patent D454,582.)

Provisional Applications

A provisional application is a patent application and must have a written description and drawings sufficient to teach the invention to one skilled in the art. A provisional application is never examined, however, and no claims are required. It is relatively inexpensive to file and provides a priority date for any application filed within a year claiming the benefit of the provisional application, as to the matter disclosed in the provisional application. Note, however, that a provisional application is automatically abandoned 1 year from its filing date. Provisional applications do not issue. A nonprovisional application must be filed within a year to receive the benefit of the filing date. Although a provisional application establishes a priority date, the 20-year term of an issued patent claiming the benefit of the provisional application begins on the filing date of the first nonprovisional application in the priority chain.

NON-U.S. PATENTS

General Information

Patents are, of course, available in other countries as well as in the United States. For example, one may apply for and obtain a patent in almost any country in the world. One may file a first application almost anywhere in the world, follow up with applications in other countries within 1 year of filing the first application, and obtain the first application's filing date as a priority date. Besides filing in individual countries, there are several regional areas in which applications can be made. These include the European Patent Convention (EPC; http://www.epo.org); the African Intellectual Property Organization (OAPI; http://www.oapi.wipo.net), the members of which are French-speaking African countries; the African Regional Industrial Property Organization (ARIPO; http://www.aripo.wipo.net), the members of which are English-speaking African countries; and the Eurasian Patent Organization (EAPO), the members of which include former republics of the Union of Soviet Socialist Republics (http://www.eapo.org).

For protection in European countries that are EPC members, an application is filed at the European Patent Office (EPO) in Munich, Germany, designating some or

all EPC member countries. Only one application needs to be filed and prosecuted to obtain coverage for any or all of the member countries. The application can be filed and prosecuted in English up to the point of issuance. A separate national patent issues for each selected country, and each patent is subject to the patent laws of the country in which it issues. A separate European Community Patent may become available in the next few years, wherein an application filed under this regime would issue as a single community-wide patent, subject ot a single jurisdiction with regard to various legal claims.

Yet another alternative is to file an application according to the Patent Cooperation Treaty (PCT; see http://www.wipo.int/pct/en). A PCT application serves as an application in each country designated in the application. The process begins when an applicant from a member country files a PCT application in a designated receiving office (such as the U.S. Patent and Trademark Office). This begins the "international stage." The application is published 18 months after the priority date. A search is performed and the search results are returned to the applicant. At the request of the applicant, a preliminary examination may be performed. This is similar to patent prosecution in the United States, although the applicant typically has just one chance to respond to a Written Opinion and to amend the claims. Thirty months (31 months for some countries) from the priority date, or 20 months in some countries if the preliminary examination has not been requested, the application must enter the *national stage* in those countries or regions in which protection is sought. Prosecution of the application then continues independently in each country or region until grant.

Filing patent applications in multiple jurisdictions can be expensive. To keep costs down, applicants typically file only in those jurisdictions where the invention is likely to be used most frequently and where meaningful enforcement can be achieved. A PCT application enables an applicant, for a relatively low cost, to delay for up to 30 months (31 months in some cases) both the designation of particular countries or regional jurisdictions and the costs of entering in those countries and regions.

Although beyond the scope of this chapter, there are foreign filing licensing requirements that must be considered before filing foreign or international (i.e., PCT) applications.

Differences Between the United States and Other Countries

Although many of the procedures and rights are similar, some differences between the United States and other nations exist. For one, in the United States (and Canada and Australia), an application may be filed up to 1 year after a publication that describes the invention. In most, if not all, other countries, such a disclosure prior to the filing of an application is an absolute bar to obtaining a patent. Another difference is that in the United States, when two inventors claim to have independently invented the invention, the patent is awarded to the first to invent (with some caveats). Of course, the precise instance of invention may be difficult to discern and to prove, leading to a complicated procedure known as interference proceedings. The rest of the world follows a "first-to-file" policy, in which a patent for an invention is awarded to the first applicant to file, regardless of who invented first.

Another difference is the term of a patent. Although patents issuing on U.S. applications filed after June 8, 1995, have a term of up to 20 years from the priority date, which is the same as foreign patents, as mentioned previously, earlier patents have a term of 17 years from issuance, and there are still many patents with this term in force.

Computer programs are also handled differently. In the United States, a computer program is patentable if it produces a "useful, concrete, and tangible result" (*State Street*,1998). In many foreign countries, however, software per se is explicitly barred from being patentable.

Another notable difference is that in the United States, the applicants must be the inventors (although they can assign their rights), whereas in other countries, the applicant may be either the inventor or the assignee. Also, in the United States, examination of a patent application by an examiner is automatic. In many other countries (e.g., Japan and South Korea), examination must be requested within some time period from the filing date. For example, in Japan, the examination must be requested within 3 years for a patent application filed on or after October 1, 2001. For applications filed prior to that date, the request must be within 7 years.

CONCLUSION

A patent does not give the owner the right to practice the invention, that is, the right to make, use, sell or offer for sale the invention in the United States, to import the invention into the United States, or to import into the United States something made by a patented process, but rather gives the inventor the right to exclude others from doing these things. That is, in return for disclosing the invention, the patentee is granted a limited term (typically 20 years from filing) in which the competition cannot use the patented device or method. The competition therefore must research and develop noninfringing alternatives, a process that could be costly, timely, and, in some cases, futile, giving the patentee a substantial advantage over the competition. Alternatively, the patentee may choose to license the patented technology and collect a royalty. For these and other reaons discussed above, patents are valuable assets for any business, large or small.

GLOSSARY

Claim The part of a patent which distinctly calls out the inventive subject matter that is protected by the patent.

Dependent claim A patent claim that incorporates by reference the limitations of another patent claim. Such a claim typically begins with language such as "The device of Claim X, further comprising" in which the limitations of Claim X are incorporated by reference.

Doctrine of claim differentiation A doctrine under which two claims of a patent are presumptively of different scope, so that if a dependent claim includes a narrowing definition of some limitation of a base claim, then the base claim is presumed to not only encompass the narrow definition, but other embodiments as well.

Doctrine of equivalents A doctrine by which, even if a device or process does not exactly match the claims of a patent, the device may still infringe the patent if the differences are insubstantial.

Element Although "element" and "limitation" are sometimes used interchangeably, the term "element" is used more frequently by the courts to refer to aspects of an alleged infringing device or method. In an infringement case, the elements of the alleged infringing device or method are compared with the claim limitations of the patent allegedly being infringed. Note, however, that 35 U.S.C. §112 refers to both an "element in a claim" (sixth paragraph) and "limitations of [a] claim" (fourth and fifth paragraphs).

Independent claim A patent claim that does not incorporate any other patent claim by reference.

Laches An equitable principle whereby a party is estopped (not allowed) from bringing a lawsuit after an unreasonable or unexplained delay which has a detrimental effect on the party being sued.

Limitation Part of a claim that defines a particular aspect of the invention. Every claim has at least one limitation, and most claims have two or more limitations.

Means plus function A particular claim limitation may be written in "means plus function" language, wherein the claim is a means or a step for performing a specified function, without reciting particular structure, material or acts, which are thus construed to be those described in the specification (and equivalents). "Means plus function" claim elements are specifically authorized by 35 U.S.C. §112, sixth paragraph.

Nonobvious One of the basic requirements in obtaining a patent is that the invention be nonobvious to one of ordinary skill in the particular art concerned, in view of the known (prior) art.

Novelty One of the basic requirements in obtaining a patent is that the invention be novel. Legally, this means that the invention must not be barred by any of the conditions stated in 35 U.S.C. §102.

Patent A grant for a fixed term that gives the holder certain rights to exclude others from practicing the patented invention as claimed. Title 35 of the United States Code provides the patent law statutes.

Patent Cooperation Treaty (PCT) A treaty under which an applicant of a member country can file a single patent application that may designate one, many, or all of the member countries. A PCT patent application is an application. The PCT does not grant patents.

Prior art The accumulated knowledge of those skilled in the particular art concerned, which can bar issuance of a patent if the claims are not novel or nonobvious in view of the prior art. Prior art that can bar issuance of a patent in the United States is defined in 35 U.S.C. §102.

Prosecution The process of obtaining a patent, from the filing of an application to the issuance of the patent (or abandonment of the application).

Prosecution laches An equitable principle in which unreasonable delay during prosecution of a patent application, together with harm to another party caused by the delay, can result in unenforceability of the issued patent.

Provisional application A patent application designated as such that is never examined but serves to provide a priority date.

Provisional rights The rights of a patent owner to a reasonable royalty for an infringing device or method, covering the period between publication of a patent application and the granting of the patent with substantially identical claims as those in the published application.

U.S. Patent and Trademark Office (USPTO) The U.S. agency authorized to grant and issue patents and to facilitate the registration of trademarks.

CROSS REFERENCES

See *Copyright Law; Legal, Social and Ethical Issues; Trademark Law.*

REFERENCES

American Inventor's Protection Act of 1999, Pub. L. No. 106–113.

Beauregard, G. M. (n.d.) *U.S. Patent No. 4,962,468.* Washington, DC: U.S. Patent and Trademark Office.

Diamond v. Chakrabarty, 447 US 303, 206 USPQ 193 (1980).

Festo Corporation v. Shoketsu Kinzoku Kogyo Kabushiki Co., Ltd., 95-1066 (Fed. Cir., 1999).

Holmes Group, Inc. v. Vornado Air Circulation Systems, Inc. (U.S. Supreme Court, 2002), (slip opinion at http://www.supremecourtus.gov/opinions/01pdf/01–408.pdf).

J. E. M. AG Supply, Inc. v. Pioneer Hi-Bred International, Inc. (99-1996) 200 F.3d 1374 (2001).

Kraft Foods, Inc. v. Int'l Trading Co., No. 99-1240, 2000 U.S. App. LEXIS 1994, 2000 WL 156556 (Fed. Cir. February 14, 2000).

Newman v. Quigg, 877 F.2d 1575 (Fed. Cir. 1989).

Platz, Axel et al., *U.S. Patent No. D453,769.* Washington, DC: U.S. Patent and Trademark Office.

Slimbach, Robert J., et al., *U.S. Patent No. D454,582.* Washington, DC: U.S. Patent and Trademark Office.

State Street Bank & Trust Co. v. Signature Financial Group, 149 F.3d 1368 (Fed. Cir. Jul. 23, 1998).

Patents, 35 U.S.C. (July 19, 1952).

Symbol Technologies, Inc. et al. v. Lemelson Medical, Education & Research Foundation, Limited Partnership, 277 F.3d 1361 (Fed. Cir. 2002).

United States Patent and Trademark Office. (2001). Guidelines for examination of design patent applications for computer-generated icons. In *Manual of Patent Examining Procedure* (8th ed.). Retrieved February 22, 2003, from http://www.uspto.gov/web/offices/pac/mpep/index.html

FURTHER READING

Books

Albert, G. P., Jr., Laff, W., &; Laff, S. (1999). *Intellectual property law in cyberspace.* Washington, DC: BNA Books.

Chisum, D. (2002). *Chisum on patents, a treatise on the law of patentability, validity and infringement.* Newark,

NJ: LexisNexis.Donner, I. H. (1999). *Patent prosecution, practice & procedure before the U.S. Patent Office* (2nd edition). Washington, DC: BNA Books.
Faber, R. C., & Landis, J. L. (1997). *Landis on mechanics of patent claim drafting* (4th ed.). New York: Practicing Law Institute.
Harmon, R. L. (2001). *Patents and the federal circuit* (5th ed.). Washington, DC: BNA Books.
Miller, A. R., & Davis, M. H. (1990). *Intellectual property—patents, trademarks and copyright in a nutshell*. St. Paul, MN: West.
Stobbs, G. A. (1995). *Software patents*. New York: Wiley.

Web Sites

American Bar Assocation (information on intellectual property): http://www.abanet.org/intelprop/comm106/106general.html
American Intellectual Property Law Assocation: http://www.aipla.org
Cornell University Law School Patent Law Overview: http://www.law.cornell.edu/topics/patent.html
Findlaw: http://www.findlaw.com
Intellectual Property Owners Association: http://www.ipo.org
Intellectual Property Today (Omega Communications): http://www.iptoday.com
IPWatchdog.com: http://www.ipwatchdog.ptcom/patent.html
Nolo: http://www.nolo.com.
PatentLawLinks.com: http://www.patentlawlinks.com
U.S. Patent and Trademark Office: www.uspto.gov
CAFC opinions may be found at
Emory University Law School Web site: http://www.law.emory.edu/fedcircuit/
Georgetown University: http:// http://www.ll.georgetown.edu/federal/judicial/cafed.cfm (Georgetown University)
The CAFC Web site: http://www.fedcir.gov
The U.S. Supreme Court Web site: http://www.supremecourtus.gov

Peer-to-Peer Systems

L. Jean Camp, *Harvard University*

CLIENTS, SERVERS, PEERS

Peer-to-peer systems (P2P) are the result of the merger of two distinct computing traditions: the scientific and the corporate. Understanding the two paths that merged to form P2P illuminates the place of P2P in the larger world of computing. Thus peer-to-peer computing when placed in historical context is both innovative and consistent with historical patterns. This larger framework assists in clarifying the characteristics of P2P systems and identifying the issues that all such systems must address by design. Recall that the core innovation of P2P is that the systems enable Wintel (Windows/Intel) desktop computers to function as full participants on the Internet, and the fundamental design requirement is coordination.

Computers began as centralized, hulking, magnificent creations. Each computer was unique and stood alone. Computers moved into the economy (beyond military uses) primarily through the marketing and design of IBM. When a mainframe was purchased from IBM it came complete. The operating systems, the programming, and (depending on the purchase size) sometimes even a technician came with the machine. Initially mainframe computers were as rare as supercomputers are today. Machines were so expensive that the users were trained to fit the machine, rather than the software being designed for the ease of the user. The machine was the center of the administrative process as well as a center of computation. The company came to the machine.

Technical innovation (the front-end processor and redesigned IBM machines) made it possible to reach multiple mainframes from many locations. Front-end processors allowed many terminals to easily attach to a single machine. Thus the first step was taken in bringing access to the user in the corporate realm. Processing power could be widely accessed through local area networks (LAN). Yet the access was through terminals with little processing power and no local storage. The processor and access remained under the administrative control of a single entity. While physical access was possible at a distance, users were still expected to learn arcane commands while working with terse and temperamental interfaces.

In parallel with the adoption of computing in the corporate world, computing and communications were spreading through the scientific and technical domains. The ARPANET (the precursor to the Internet) was first implemented in order to share concentrated processing power in scientific pursuits. Thus the LAN was developing in the corporate realm while the wide area network (WAN) was developing in the world of science.

Before the diffusion of desktop machines, there were so-called microcomputers on the desktops in laboratories across the nation. These microcomputers were far more powerful than concurrent desktop machines. (Currently microcomputers and desktop computers have converged because of the increase in affordable processing power.) Here again the user learned to communicate based on the capacities of the machine. These users tended to embrace complexity; thus they altered, leveraged, and expanded the computers.

Because microcomputers evolved in the academic, scientific, and technical realm the users were assumed to be capable managers. Administration of the machines was the responsibility of the individual users. Software developed to address the problems of sharing files and resources assumed active management by end users. The early UNIX world was characterized by a machine being both a provider and a consumer of services, both overseen with a technically savvy owner/manager.

The Internet came from the realm of the UNIX world, which evolved independently of the desktop realm. Comparing the trajectories of e-mail in the two realms is illustrative. On the desktop, e-mail evolved in proprietary environments where the ability to send mail was limited to those in the same administrative domain. Mail could be centrally stored and was accessed by those with access rights provided by a central administrative body. In contrast, in UNIX environments, the diffusion of e-mail was enabled by each machine having its own mail server. For example, addresses might be michelle@smith.research.science.edu in one environment as opposed to john_brown@vericorp.web in the other. (Of course early corporate mail services did not use domain names, but this fiction simplifies the example.) In the first

case Michelle has a mail server on her own UNIX box; in the second John Brown has a mail client on his machine that connects to the shared mail server being run for Vericorp. Of course, now the distinct approaches to e-mail have converged. Today users have servers that provide their mail, and access mail from a variety of devices (as with early corporate environments). E-mail can be sent across administrative domains (as with early scientific environments). Yet the paths to this common endpoint were very different with respect to user autonomy and assumptions about machine abilities.

The Internet and UNIX worlds evolved with a set of services assuming all computers were contributing resources as well as using them. In contrast, the Wintel world developed services where each user had corresponding clients to reach networked services, with the assumption that connections were within a company. Corporate services are and were provided by specialized powerful PCs called (aptly) servers. Distinct servers offer distinct services with one service per machine or multiple services running from a single server. In terms of networking, most PCs either used simple clients, acted as servers, or connected to no other machines.

Despite the continuation of institutional barriers that prevented early adoption of cross-corporate WANs, the revolutionary impact of the desktop included fundamentally altering the administration, control, and use of computing power. Standalone computers offered each user significant processing ability and local storage space. Once the computer was purchased, the allocation of disk space and processing power were under the practical discretion of the individual owner. Besides the predictable results, for example the creation of games for the personal computer, this required a change in the administration of computers. It became necessary to coordinate software upgrades, computing policies, and security policies across an entire organization instead of implementing the policies in a single machine. The difficulty in enforcing security policies and reaping the advantages of distributed computing continues, as the failures of virus protection software and proliferation of vulnerabilities illustrates.

Computing on the desktop provides processing to all users, offers flexibility in terms of upgrading processing power, reduces the cost of processing power, and enables geographically distributed processing to reduce communications requirements. Local processing made spreadsheets, "desktop" publishing, and customized presentations feasible. The desktop computer offered sufficient power that software could increasingly be made to fit the users, rather than requiring users to speak the language of the machines.

There were costs to decentralization. The nexus of control diffused from a single administered center to across the organization. The autonomy of desktop users increases the difficulty of sharing and cooperation. As processing power at the endpoints became increasing affordable, institutions were forced to make increasing investments in managing the resulting complexity and autonomy of users.

Sharing files and processing power is intrinsically more difficult in a distributed environment. When all disk space is on a single machine, files can be shared simply by altering the access restrictions. File sharing on distributed computers so often requires taking a physical copy by hand from one to another that there is a phrase for this action: sneakernet. File sharing is currently so primitive that it is common to e-mail files as attachments between authors, even within a single administrative domain. Thus currently the most commonly used file-sharing technology remains unchanged from the include statements dating from the sendmail on the UNIX boxes of the 1980s.

The creation of the desktop is an amazing feat, but excluding those few places that have completely integrated their file systems (such as Carnegie Mellon which uses the Andrew File System) it became more difficult to share files, and nearly impossible to share processing power. As processing and disk space become increasingly affordable, cooperation and administration became increasingly difficult.

One mechanism to control the complexity of administration and coordination across distributed desktops is a client–server architecture. Clients are distributed to every desktop machine. A specific machine is designated as a server. Usually the server has more processing power and higher connectivity than the client machines. Clients are multipurpose, according to the needs of a specific individual or set of users. Servers have either one or few purposes; for example, there are mail servers, Web servers, and file servers. While these functions may be combined on a single machine, such a machine will not run single-user applications such as spreadsheet or presentation software. Servers provide specific resources or services to clients on machines. Clients are multipurpose machines that make specific requests to single-purpose servers. Servers allow for files and processing to be shared in a network of desktop machines by reintroducing some measure of concentration. Recall that peers both request and provide services. Peer machines are multipurpose machines that may also be running multiple clients and local processes. For example, a machine running Kazaa is also likely to run a Web browser, a mail client, and a MP3 player. Because P2P software includes elements of a client and a server, it is sometimes called a *servlet*.

Peer-to-peer technology expands file- and power-sharing capacities. Without P2P, the vast increase in processing and storage power on the less-predictable and more widely distributed network cannot be utilized. Although the turn of the century sees P2P as a radical mechanism used by young people to share illegal copies, the fundamental technologies of knowledge sharing as embedded in P2P are badly needed within government and corporate domains.

The essence of P2P systems is the coordination of those with fewer, uncertain resources. Enabling any party to contribute means removing requirements for bandwidth and domain name consistency. The relaxation of these requirements for contributors increases the pool of possible contributors by order of magnitude. In previous systems sharing was enabled by the certainty provided by the technical expertise of the user (in science) or administrative support and control (in the corporation). P2P software makes end-user cooperation feasible for all by simplification of the user interface.

PCs have gained power dramatically, yet most of that power remains unused. While any state-of-the-art PC purchased in the past five years has the power to be a Web server, few have the software installed. Despite the affordable migration to the desktop, there remained a critical need to provide coordinated repositories of services and information.

P2P networking offers the affordability, flexibility, and efficiency of shared storage and processing offered by centralized computing in a distributed environment. In order to effectively leverage the strengths of distributed coordination P2P systems must address reliability, security, search, navigation, and load balancing.

P2P systems enable the sharing of distributed disk space and processing power in a desktop environment. P2P brings desktop Wintel machines into the Internet as full participants.

Peer-to-peer systems are not the only trend in the network. Although some advocate an increasingly stupid network and others an increasingly intelligent network, what is likely is an increasingly heterogeneous network.

FUNCTIONS OF P2P SYSTEMS

There are three fundamental resources on the network: processing power, storage capacity, and communications capacity. Peer-to-peer systems function to share processing power and storage capacity. Different systems address communications capacity in different ways, but each attempts to connect a request and a resource in the most efficient manner possible.

There are systems that allow end users to share files and file and groupware processing power. Yet none of these systems are as effective as peer to peer systems. However, all of these systems solve the same problems as P2P systems do: naming, coordination, and trust.

Mass Storage

As the sheer amount of digitized information increases, the need for distributed storage and search increases as well. Some P2P systems enable sharing of material on distributed machines. These systems include Kazaa, Publius, Free Haven, and Gnutella. (Limewire and Morpheus are Gnutella clients.)

The Web enables publication and sharing of disk space. The design goal of the Web was to enable sharing of documents across platforms and machines within the high-energy physics community. When accessing a Web page a user requests material on the server. The Web enables sharing, but does not implement searching and depends on DNS for naming. As originally designed the Web was a P2P technology. The creation of the browser at the University of Illinois Urbana–Champaign opened the Web to millions by providing an easy-to-use graphical interface. Yet the dependence of the Web on the DNS prevents the majority of users from publishing on the Web. Note the distinction between the name space, the structure, and the server as constraints.

The design of the hypertext transport protocol (HTTP) does not prevent publication by an average user. The server software is not particularly complex. In fact, the server software is built into Macintosh OS X. The constraints from the DNS prevent widespread publication on the Web. Despite the limits on the namespace, the Web is the most powerful mechanism used today for sharing content. The Web allows users to share files of arbitrary types using arbitrary protocols. Napster enabled the sharing of music. Morpheus enables the sharing of files without constraining the size. Yet neither of these allows the introduction of a new protocol in the manner of HTTP.

The Web was built in response to the failures of distributed file systems. Distributed files systems include the network file system and the Andrew file system, and are related to groupware. Lotus Notes is an example of popular groupware. Each of these systems shares the same critical failure—institutional investment and administrative coordination are required.

Massively Parallel Computing

In addition to sharing storage P2P systems can also share processing power. Examples of systems that share processing power are Kazaa and SETI@home.

There are mechanisms other than P2P systems to share processing power. Such systems run only on UNIX variants, depend on domain names, are client–server, or are designed for use only within a single administrative domain. Metacomputing and clustering are two approaches to sharing processing power. Despite the difference in platform, organization, and security, the naming and organization questions are similar in clusters and peering systems.

Clustering systems are a more modern development. Clustering software enables discrete machines to run as a single machine. Beowulf came from NASA in 1993 (Wulf et al., 1995). The first Beowulf Cluster had 16 nodes (or computers) and the Intel 80486 platform. (Arguably this was more than a money-saving innovation as it was a pivotal moment in the fundamental paradigmatic change in the approach to supercomputing reflected in P2P.) DAISy (Distributed Array of Inexpensive Systems) from Sandia was an early provider of a similar functionality. Yet these systems are descended from the UNIX branch of the network tree. Each of these systems are built to harness the power of systems running Linux, as opposed to running on systems loaded with the Windows operating system. (Linux systems are built to be peers, as each distribution includes, for example, a Web server and browser software as well as e-mail servers and clients.)

Clustering systems include naming and distribution mechanisms. Recall Beowulf, an architecture enabling a supercomputer to be built out of a cluster of Linux machines. In Beowulf, the machines are not intended to be desktop machines. Rather the purpose of the machines is to run the software distributed by the Beowulf tree in as fast a manner as possible. Beowulf is not a single servlet. Beowulf requires many elements, including message-passing software and cluster management software, and is used for software designed for parallel systems. Beowulf enables the same result as that provided by a P2P processor-sharing system: the ability to construct a supercomputer for a fraction of the price. Yet Beowulf assumes the clusters are built of single-purpose machines within a single administrative domain.

EXAMPLES OF P2P SYSTEMS

In this section the general principles described above are discussed with respect to each system. For each system design goals and organization (including centralization) are discussed. Mechanisms of trust and accountability in each system are described.

Given the existence of a central server there are some categorizations that place SETI@home and Napster outside the set of P2P systems. They are included here for two reasons. First for theoretical reasons, both of these systems are P2P in that they have their own name spaces and utilize heterogeneous systems across administrative domains in cooperative resource sharing (Oram, 2001). Second, any definition that is so constrained as to reject the two systems that essentially began the P2P revolution may be theoretically interesting but are clearly flawed.

P2P systems are characterized by utilization of desktop machines that lack domain names, experience intermittent connectivity, have variable connection speeds when connected, and possibly have variable connection points (for laptops, or users with backup ISPs).

Napster

Napster began as a protocol, evolved to a Web site, became a business with an advertising-driven value of millions, and is now a wholly owned subsidy of Bertelsmann entertainment. Yet the initial design goal was neither to challenge copyright law nor to create a business; the original goal was to enable fans to swap music in an organized manner. Before Napster there were many Web sites, ftp sites, and chat areas devoted to locating and exchanging music files in the MPEG3 format; Napster, however, simplified the location and sharing processes. The goal of Napster was to allow anyone to offer files to others. Thus the clients were servers, and therefore Napster became the first widely known P2P system.

Before Napster, sharing music required a server. This required a domain name and specialized file transfer software or streaming software. The Napster client also allowed users to become servers, and thus peers. The central Napster site coordinated the peers by providing a basic string-matching search and the file location. As peers connected to Napster to search, the peers also identified the set of songs available for download.

After Napster the client software was installed on the peer machine and contacted http://www.napster.com, Napster the protocol then assigned a name to the machine. As the peer began to collect files it might connect from different domains and different IP addresses. Yet whether the machine was connected at home or at work Napster could recognize the machine by its Napster moniker.

Thus Napster solved the search problem by centralization and the problem of naming by assignment of names distinct from domain names.

When a peer sought to find a file, the peer first searched the list of machines likely to have the file at the central Napster archive. Then the requesting peer selected the most desirable providing peer, based on location, reputation, or some other dimension. The connection for obtaining the file was made from the requesting peer to the providing peer, with no further interaction with the central server. After the initial connection the peer downloaded the connection from the chosen source. The chosen source by default also provided a listing of other songs selected by that source.

Accountability issues in Napster are fairly simple. Napster provided a single source for the client; therefore downloading the peer-to-peer software needed to join the network was not an issue of trust. Of course, the Napster Web site itself must be secure. Napster had been subject to attacks by people uploading garbage files but not by people upload malicious files.

In terms of trust, each user downloaded from another peer who was part of the same fan community. Grateful Dead fans share music as do followers of the Dave Matthews Band. Each group of fans shared music within their communities. It is reasonable to assert that Napster was a set of musical communities, as opposed to a single community of users.

Kazaa

Kazaa is a P2P system optimized for downloads of large files. Unlike the hobbyist or scientific basis of many P2P systems, the widely installed Kazaa software has always been a business first. Kazaa is downloaded by users presumable for the access to music and, in particular, large video files on remote machines. Kazaa was created by a team of Dutch programmers and then sold to Sharman Networks. In 2002 Kazaa was downloaded by more than 120 million users. Kazaa has always sold advertising, charging to access the customers' attention span. Kazaa has decentralized search and file distribution.

Kazaa also installs up to four types of additional software in order to enhance its revenue stream. First, and most importantly for Kazaa, the software installs an ad server. Kazaa's business model depends on advertiser revenue. Kazaa installs media servers to enable high-quality graphics in its advertising.

Second, Kazaa installs software to use processing resources on the users' machines. Sharman Networks has teamed with Brilliant Networks to develop software that enables processing power to be shared. With a centralized command the Brilliant Software owners can utilize the processing power of all Kazaa users. As of the close of 2002, the system is not being used to resell processing power. Company statements suggest it is being used to allow machines to serve ads to others (Borland, 2002).

Third, Kazaa installs media servers that allows complex video advertisements.

Fourth, Kazaa alters affiliate programs. Many companies get a percentage of purchases. Affiliate programs are used by businesses, not-for-profits, and individuals. Kazaa intercepts affiliate messages and alters the flow of revenue to Kazaa.

In some versions, Kazaa includes a shop-bot, which compares prices while the user shops using a browser. The shop-bot identifies sites with better prices when the user seeks an identifiable good.

Kazaa also offers New.net—an alternative domain name root. By enabling an alternative root New.net allows

users to choose domains names other than the original top-level domain names and allows domain name registrants to maintain their own privacy. (The governing body of the original top-level domain names increasingly requires identifying information whenever a domain name is purchased. Anonymous domain names, and thus anonymous speech, are increasingly disallowed in the top-level domains controlled by ICANN.)

In terms of trust the user must trust Kazaa and trust other users.

In order to encourage users to cooperate Kazaa has a *participation level*. According to a competitor (K-lite) the participation level measures the ratio of downloads to uploads. Depending on this ratio the speed of downloads is altered. A user who offers popular content is allowed higher access speeds than users who download but do not upload.

According to Kazaa the participation level only matters if there is competition for a file. If two or more users seek to access a file then the user with the higher participation level has priority. According to K-lite there are controls on download speeds for all access attempts.

Besides offering uploads, another way to increase a participation level is to increase the detail of metadata available about a file. Integrity is a measure of the quality of the descriptors of the data. Metadata describes the content, including identifying infected or bogus files by rating them as "D." "Integrity level" is another trust mechanism implemented with Kazaa. This means that the descriptors may be good regardless of the quality of the data. Other descriptions include content and technical quality.

Kazaa implements mechanisms to enables users to trust each other and trust the content downloaded. Kazaa does not implement technical mechanisms to encourage the user to trust Kazaa itself. Kazaa offers stated privacy policies for all the downloaded software. However, the difference between the descriptions of participation level at Kazaa and K-lite suggests that there is distrust. In addition, the prominent declaration on Kazaa's site that there is no spyware in October 2002 in Kazaa suggests that there is indeed concern. This declaration directly contradicts media reports and the description of competitors describing the installation of spyware by Kazaa. (See Lemos, 2002.)

Search for Intelligent Life in the Universe

SETI@home distributes radio signals from the deep space telescope to home users so that they might assist in the search for intelligent life. The Arecibo telescope sweeps the sky collecting 35 Gbyte of data per day.

To take part in this search, each user first downloads the software for home machine use. After the download the user contacts the SETI@home central server to register as a user and obtain data for analysis. Constantly connected PCs and rarely connected machines can both participate.

There are other projects that search for intelligent life via electromagnetic signals. Other programs are limited by the available computing power. SETI@home allows users to change the nature of the search, enabling examination of data for the weakest signals.

SETI@home is indeed centralized. There are two core elements of the project—the space telescope at Arecibo and the peer-to-peer analysis system. Each user is allocated data and implements analysis using the SETI software. After the analysis the user also receives credit for having contributed to the project.

SETI tackles the problem of dynamic naming by giving each machine a time to connect, and a place to connect. The current IP address of the peer participant is recorded in the coordinating database.

SETI@home is P2P because it utilizes the processing power of many desktops, and uses its own naming scheme in order to do so. The amount of data examined by SETI@home is stunning, and far exceeds the processing capacity of any system when the analysis is done on dedicated machines. SETI is running 25% faster in terms of floating point operations per second at 0.4% of the cost of the supercomputer at Sandia National Laboratories. (The cost ratio is 0.0004.) SETI@home has been downloaded to more than 100 countries. In July 2002 there were updates to the SETI software in Bulgarian, Farsi, and Hebrew.

The software performs Fourier transforms—a transformation of frequency data into time data. The reason time data are interesting is that a long constant signal is not expected to be part of the background noise created by the various forces of the universe. Finding a signal that is interesting in the time domain is indicative of intelligent life.

The client software can be downloaded only from SETI@home in order to make certain that the scientific integrity of code is maintained. If different assumptions or granularity are used in different Fourier analyses, the results cannot be reliably compared with other results using original assumptions. Thus even apparently helpful changes to the code may not, in fact, be an improvement.

SETI@home provides trustworthy processing by sending out data to different machines. This addresses both machine failures and malicious attacks. SETI@home has already seen individuals altering data to create false positives. SETI@home sends data to at least two distinct machines, randomly chosen, and compares the results. Note that this cuts the effective processing rate in half, yielding a cost/processing ratio of 0.002 as opposed to a 0.004. However, the cost per processing operation remains three orders of magnitude lower for SETI@home than for a supercomputer.

SETI@home has also had to digitally sign results to ensure that participants do not send in results multiple times for credit within the SETI@home accounting system. (Since there is no material reward for having a high rating the existence of cheating of this type came as a surprise to the organizers.) SETI@home can provide a monotonically increasing reputation because the reputation is the reward for participation. In addition to having contributions listed from an individual or a group, SETI@home lists those who find any promising anomalies by name.

Gnutella

Gnutella was developed as an explicit response to the legal problems of Napster (von Lohmann, 2001). The developers of Gnutella believed that the actions labeled as theft by the owners of copyrights were in fact sharing.

Philosophical and economic arguments (qualitative and quantitative) that Napster encouraged the purchase of compact discs have been made, e.g., Pahfl (2001). Some argue that the sharing of songs on Napster was more advertisement than substitute for a purchased work. The creators of Gnutella had observed the expansion of rights of trademark holders and the ability of censors to use copyright law to prevent critical speech. (The Church of Scientology has had particular success in this legal strategy.)

Based on concepts of fair use and ideological commitments to sharing, Gnutella enables sharing of various types of files. Gnutella allows users to share their disk space for storage and search by integrating the search into the client.

Gnutella searches works on the basis of local broadcasts. Each peer is connected to *n* other peers in a search pattern, and so on down the line. If a peer receives a query that it can answer, it responds affirmatively. If the peer does not have the requested content then the receiving peer resends the query to its immediate peers. Because Gnutella is built in this modular fashion, shutting down a single peer will not prevent sharing. Gnutella applications can exist in a large networked tree or as independent cells.

The broadcast model of searching is considered to be a weakness with respect to the ability to scale (Ritter, 2002). However, Gnutella's search technique allows local cells to survive without broader connections and implements a very through search. Gnutella enables scaling through segmenting the network. Gnutella creates a small world network, where there is a network of closely connected nodes and few connections between the networks. The design is based on the six-degrees-of-separation concept (familiar to some as the Kevin Bacon game).

In Gnutella the searches are made anonymous, yet downloads are not. Thus there is the assumption that the server contacted by a requester will not log the request. Yet this assumption has not held up in practice. Gnutella requires that requestors trust providers. The trust assumption has been used to entrap criminals. In particular, some users work to defeat the use of Gnutella to trade child pornography. By using a tool to generate fake file names combining explicit words and young ages and logging the file, it is fairly simple to post deceptively named files and create a "Wall of Shame," publicly showing the IP address of those who request the files. In this case the lack of anonymity enabled social accountability. Of course, the same techniques can be used to bait those interested in files about Chinese Democracy or open source software; yet in 2000 there was no record of the practice. The example of the Wall of Shame illustrates the complexity of the issue of accountability in distributed anonymous systems.

Limewire and Morpheus

Limewire and Morpheus are implementations of the Gnutella protocol. Currently Limewire is the most popular as a Macintosh servlet while Morpheus dominates the Wintel world. Morpheus is also available for the Macintosh platform. Limewire is written in Java and is available for all platforms. (As of October 2002, Limewire is available for 12 platforms.) The source of Limewire is available, theoretically preventing some of the revenue-capturing methods of Kazaa. (Of course, Limewire could make the same arrangement with New.net, as described below.)

Limewire offers a version without advertisements for $9.50 and with advertisements for free. (Note that Opera uses the same strategy.) The version with ads installs ClickTillUWin.com—a bit of adware that pops windows up as long as the program is active.

Limewire has developed a two-tier network. There are coordinating peers (called ultrapeers) who assist in searching and organizing downloads. These are used to optimize the system for all users. The standard peers connect to one or two ultrapeers. The ultrapeers do not host the files, but rather organize downloads. Each ultrapeer is associated with a subnet, and the ultrapeers are themselves tightly connected.

In order to increase the speed of downloads and distribute the load on peer-providing files Limewire uses swarming transfers. Swarm downloading entails downloading different elements of files available on multiple low-bandwidth connections to obtain the equivalent service of a single broadband connection. Swarming prevents concentration of downloads from a single server as well. Essentially swarm downloading provides decentralized load balancing.

Limewire implements accountability by allowing a source to obtain information about the number of files shared by a requester. If a peer requesting a file does not offer many files to others, the peer receiving the request may automatically refuse to share any files with the requester.

Morpheus similarly offers source code availability. Morpheus bundles its code with adware, as with Limeware. Morpheus also installs software to resell disk space and CPU cycles. Early on Morpheus redirected affiliation programs to Morpheus; however, this appears to have ended in later versions.

Mojo Nation

Mojo Nation implements a shared P2P system to enable reliable publishing at minimal cost. Mojo Nation solves the problem of availability and uses microcurrency to address the problem of persistence. Mojo Nation is designed to prevent free riding. Mojo Nation implements a micropayment system and a reputation system using a pseudocurrency called Mojo. Mojo is not convertible.

Mojo Nation software combines the following functions: search, download, search relay, and publishing content. Search relay is used to support the searches of other users.

The Mojo Nation is designed to provide reliability. Mojo Nation provides reliability by using a swarm download. Any participant in Mojo Nation is pseudonymous. Mojo identities are public keys (RSA) generated by the user's own computer. The pseudonym can then be moved with the machine, and is associated with its own Mojo balance.

Mojo Nation is not optimized for a particular type of file. Swarm downloading enables downloads of large files, while small files present no particular problem. Examples of large files include video, while MP3 files are smaller and more easily managed.

Mojo is designed neither to enable (as with Publius) nor to disable (as with Free Haven [Sniffen, 2000]) file deletion. Files may be published individually or as an explicit collection. Publication of collections enables the creation of user lists and collaborative filtering, as with Napster and Amazon.

Mojo peers can search for content, download content, support others' searches, offer content, or relay search information. The relay function enables users who are behind firewalls or otherwise blocked to continue to use Mojo Nation.

Searching and downloading cost Mojo while supporting others' searches and providing content earns Mojo. Each downloaded peer software package begins with some Mojo, so pseudo-spoofing assaults on the system are possible. Pseudo-spoofing is the creation of many accounts in order to built the account or reputation of some single account.

Distributed.net

Distributed.net is a volunteer cooperative association of individuals interested in supporting distributed computing for scientific and social gain. Distributed.net is a not-for-profit that addresses computational challenges for the public good.

One distributed.net interest is testing security mechanisms for the Internet. In particular much of the security on the Internet is based on public key encryption. The basis for public key encryption is that there exists mathematical functions that are one way with a trap door. A one-way function is one that is easy to do but hard to undo. Physical examples of one-way functions abound: words cannot be unsaid, a broken egg cannot be repaired, and a thousand tacks cannot easily be placed back in a tube. A trap door is a mathematical secret (the cryptographic key) that makes the one-way function possible to undo. While there is no trap door for eggs, replacing tacks can be made simple with a dustpan and well-matched funnel.

Distributed.net also works to examine the strength of algorithms where both users share one key. In these algorithms the information-hiding system is attacked by trying every possible key. By giving many users different keys to guess, any message can be attacked more effectively.

Distributed.net tests the difficulty of breaking the security of a most widely used encryption algorithm. The difficulty of this feat is of interest to militaries, businesses, and privacy advocates because all three groups share an interest in protecting information.

Distributed.net is an increasingly large and valuable shared resource. Distributed.net seeks computational challenges for scientific advancement. Future possible challenges include working on DNA coding or examining public health data for possible correlations of disease outbreaks or symptoms.

P2P IN BUSINESS

Corporate P2P systems seek to solve two primary problems—real-time collaboration and knowledge management.

For many users real-time collaboration currently is not feasible, especially if the collaboration crosses administrative domains. The problems of corporate networks are the problems of sharing information across domains, and (with increasingly mobile devices) identifying trusted devices and users. For many users, collaboration is implemented via attaching documents to unencrypted mail and hoping that no one is watching. While P2P is decried as being designed for theft, in fact the creation of scalable namespaces and trust networks integrated with collaborative tools is valuable for every group from families to enterprises.

Peer-to-peer systems in the corporate environment must solve the same problems as noncommercial P2P systems: disparate machines, distant locations, participants with widely different capacities, and a lack of single namespace that covers all machines. Yet in a business domain there is a central authority or chain of command that determines (and delegates) the relative authority of participants. Therefore reputation systems in P2P systems can be replaced with more formal security systems based on authentication of users. Both classes of systems offer their own namespaces.

Yet even with P2P technology there is a culture of proprietary and the practice of closed code. Therefore the descriptions here necessarily lack the detail provided for the more academic or open networks described above.

Groove

Groove, founded in October 1997, is a commercial application of P2P technology for solving the chronic problem of institutional knowledge management. The core Groove team includes Ray Ozzie, the creator of Lotus Notes. (Notes is a server-based product for sharing data and workspaces.) Groove is also of interest because it has been embraced as a standard by Microsoft. Groove is P2P in that it allows users to share material on their own machines and create a new namespace for users and files to allow this to happen.

Groove shares with Lotus Notes the concepts of accounts, identities, and shared-space membership, and conceptually expands to include presence and contacts. Unlike Lotus Notes, the participants in a Groove workspace need not share administrative access to a single server or even have a Notes client to share content.

In contrast to Lotus, Groove utilizes the capacities of users' desktops as opposed to requiring that users place their shared documents in a remote workspace. When documents are updated on the users' workspace, they are seamlessly shared.

Groove is a package of software that includes instant messaging, e-mail, document sharing, and real-time collaboration, sometimes called shared whiteboards. As Groove is tightly integrated with the Microsoft Office Suite an Office document can be shared within Groove with simultaneous instant messaging without switching between applications or requiring additional namespaces.

Groove is emphatically not a single sign-on project. Groove allows users to create multiple roles: employee, supervisor, mom, work, wife, or PTA president. Each role has its own "identity" and reputation within the sphere of its identity. This also allows users and communities to

create their own namespaces without threatening (by the temptation to leverage) corporate namespaces.

In summary Groove integrates P2P sharing of files, Microsoft Office document revision tools, e-mail, and chat in a single software package. (Given the nature of the software and the integration of other tools Groove cannot be classified as an application or communications software alone.)

Tenix

Tenix has an explicit P2P basis. Tenix offers to bring P2P to the corporate environment by adding a secure namespace and offering the option of storing files in a central location. Tenix uses a public key infrastructure (PKI) to allow corporations to verify users and create powers of delegation.

By using a public key hierarchy Tenix creates a virtual P2P network where users can identify themselves within the cryptographically secure namespace. A review of the PKI section will illustrate that one way to implement public key systems is to create a set of trusted roots. These roots then use cryptographic credentials to allow users in the same PKI to verify themselves. A Tenix identity is a single-user identity, in contrast to Groove where a Groove identity is a role. (For example, Jean Camp is a single identity while Professor Camp and Aunt Jean are distinct professional and family roles.)

Tenix creates supernodes in the same conceptual model as Kazaa.

Group membership, access control, and version information are stored on a central server. Tenix can be installed with an ultrapeer to coordinate naming and resource location. Alternatively coordination can be provided by Tenix so that the organization can choose to outsource and still leverage P2P systems.

Tenix is P2P in that it enables users to share resources, but it can be installed so that the P2P options are not fully utilized and the system a closed server architecture.

CONCLUSION

There are significant research issues with respect to digital networked information, including problems of naming, searching, organizing, and trusting information. Because peer-to-peer systems required downloading and installing code as well as providing others with access to the user's machine, the problem of trust is particularly acute. The vast majority of users of peer-to-peer systems are individuals who lack the expertise to examine code even when the source code can be downloaded and read.

Peer-to-peer systems currently are at the same state as the Web was in 1995. It is seen as an outlaw or marginal technology. As with the Web, open source, and the Internet itself the future of peer to peer is both in the community and in the enterprise.

Peer-to-peer systems solve (with varying degrees of success) the problem of sharing data in a heterogeneous network. Just as no company is now without an intranet using Web technologies, in a decade no large enterprise will be without technology that builds on today's peer-to-peer systems.

Peer-to-peer systems bring the naïve user and the Wintel user onto the Internet as full participants. By vastly simplifying the distribution of files, processing power, and search capacity peer-to-peer systems offer the ability to solve coordination problems of digital connectivity.

Peer-to-peer software is currently a topic of hot debate. The owners of high-value commodity content believe themselves to be losing revenue to the users of peer-to-peer systems. In theory all downloaded music may be lost revenue. An equally strong theoretical argument is that peer-to-peer systems now serve as a mechanism for advertising, like radio play, so that music popular on peer-to-peer networks is music that will be widely purchased.

There are strong lobbying efforts to prohibit peer to peer software. Some ISPs prohibit peer to peer software by technical and policy means.

There is debate within as well as about the peering community. By bundling software for ads, peer-to-peer systems are creating innovative business models or alternatively committing crimes against users. In one case the reselling of processing power by the software creators is seen as an innovative way to support the peer-to-peer network. From the other perspective the peers bring the value to the community and bundled software illegitimately exploits that value. Thus some decry the bundled software installed with P2P code as parasitic or spyware. Installing advertising, software that records user actions, or software that redirects affiliate programs is seen by users as a violation of the implied agreement. (The implied contract is that users share their own content and in return obtain content provided by others.)

Press coverage of peer-to-peer systems today is not unlike press coverage of early wireless users at the turn of the last century, both admiring and concerned about radical masters of frightening technology bent on a revolution. In a decade or so, peer-to-peer systems within the enterprise will be as frightening and revolutionary as radio is today. Yet without breakthroughs in the understanding of trust in the network, peer to peer across administrative domains may founder on the problems of trust and thus become, like Usenet and gopher, footnotes for the historical scholar.

ACKNOWLEDGMENT

This work was supported by the National Science Foundation under Grant No. 9985433 and a grant from the East Asia Center. Any opinions, findings, and conclusions or recommendations expressed are those of the author and do not necessarily reflect the views of the National Science Foundation.

GLOSSARY

Cluster A set of machines whose collective resources appear to the user as a single resource.

Consistency Information or system state shared by multiple parties.

Domain name A mnemonic for locating computers.

Domain name system The technical and political process of assigning, using, and coordinating domain names.

Front-end processor The first IBM front-end processor enabled users to access multiple mainframe computers from a single terminal.

Node A machine (sometimes a set of machines) that has a single specific name.

Metadata Data about data, e.g., location, subject, or value of data.

Persistent Information or state that remains reliably available over time despite changes in the underlying network.

Reliable The maintaining of a system's performance as a whole despite localized flaws or errors; e.g., file recovery despite disk errors, file transmission despite partial network failure; alternatively a system (particularly a storage system) in which all failures are recoverable, so that there is never long-term loss.

Scalable The idea that a system that works within a small domain or for a small number of units will also work for a number of units orders of magnitude larger.

Swarm A download of the same file from multiple sources.

Servlet Software-integrating elements of client and server software.

Top-level domain name The element of the domain name that identifies the class and not the specific network; early examples: edu and org.

Trusted The state of a component if the user is worse off should the component fail, the component allows actions that are otherwise impossible, i.e., it is enabling, and there is no way to obtain the desired improvement without accepting the risk of failure. (Note that reliable systems do not require trust of a specific component.)

UNIX A family of operating system used by minicomputers and servers, and increasingly on desktops; Linux is a part of the UNIX family tree.

Virus A program fragment that attaches to a complete program in order to damage the program, files on the same machine, and/or infect other programs.

CROSS REFERENCES

See *Client/Server Computing; Middleware; Web Services*.

REFERENCES

Borland, J. (2002, April 1). Stealth P2P network hides inside Kazaa. *CNET Tech News*. Retrieved from http://news.com.com/2100-1023-873181.html

Camp, L. J. (2001). *Trust and risk in Internet commerce*. Cambridge, MA: MIT Press.

Lemos, R. (2002, June 6). AIM+ creators delete "spyware" feature. *CNET Tech News*. Retrieved from http://news.com.com/2100-1040-933576.html

Oram, A. (Ed.) (2001). *Peer-to-peer harnessing the power of disruptive technologies*. Cambridge, MA: O'Reilly.

Pahfl, M. (2001). Giving music away to make money: Independent musicians on the Internet. *First Monday, 6*(8). Retrieved from http://www.firstmonday.org/issues/issue6_8/pfahl/index.html

Ritter, J. (2001, February). Why Gnutella can't scale. No, really. Retrieved from http://www.darkridge.com/~jpr5/doc/gnutella.html

Sniffen, B. (2000). *Trust economies in the Free Haven Project* (MIT technical report). Cambridge, MA: Massachusetts Institute of Technology.

von Lohmann, F. (2001). Peer to peer sharing and copyright law after Napster. Retrieved from http://www.eff.org/Intellectual_property/Audio/Napster/20010309_p2p_exec_sum.html

Wulf, W. A., Wang, C., & Kienzle, D. (1995, August). A new model of security for distributed systems (Computer Science Technical Report CS-95-34). University of Virginia.

FURTHER READING

Online

Beowulf, http://www.beowulf.org/

Free Haven, http://www.freehaven.net

Free Network Project, http://freenet.sourceforge.net

Gnutella, http://www.gnutella.org

Kazaa, http://www.kazaa.com

Kazaa Lite, http://www.k-lite.tk, a version of Kazaa without "spyware" or limits on download speed.

Mojo Nation, http://sourceforge.net/projects/mojonation

Napster Protocol, http://opennap.sourceforge.net/napster.txt

SETI@home, http://setiathome.ssl.berkeley.edu

Perl

David Stotts, *University of North Carolina at Chapel Hill*

INTRODUCTION

From its introduction to the programming community in 1987, Perl has become today one of the most widely known and used programming languages. Designed by Larry Wall, and originally thought of as a natural enhancement for the popular *csh* shell script notation of Unix, Perl was at first primarily used for text manipulation. Its maturity in the early 1990s coincided with the rise of the Web, and it rapidly became the most popular programming language for HTML form processing and other Web development as well.

Perl has been called a "Swiss Army chainsaw" for its plethora of features coupled with its considerable programming power and flexibility. The common phrase among hardened Perl programmers is "there's more than one way to do it." Most programming goals can be achieved in Perl in at least three ways, depending on which language features and techniques the programmer prefers to use. It is not uncommon for an experienced Perl programmer to reach for the manual when reading code written by another programmer. Perl has also been called "duct tape for the Web," emphasizing its utility for producing applications, Web sites, and general program fixes for a wide variety of problems and domains.

In this chapter we give a brief history of Perl, including major events preceding Perl that set the historical stage for it. We provide an overview of the language, including example code to show how its features are used in practice. We discuss Web site programming in Perl using the CGI (Common Gateway Interface) standard and show several database interface methods in Perl. We discuss the community of programmers that has grown up around Perl and conclude with a presentation of several technologies that are logical follow-ons to Perl.

A BRIEF HISTORY OF PERL

Perl grew out of the Unix programming community. Though it did not formally appear until the late 1980s, the technical components and motivations for Perl were developed in the two decades prior to that. Here are the main events in the "genealogy" of Perl:

1969 Unix is created at Bell Labs
1977 awk is invented by Aho, Weinberger, and Kernighan
1978 "sh" shell is developed for Unix
1987 Perl is created
1995 (March) Perl 5.001 released, the most recent major version;
as of this writing, Perl version 5.8.0 is the newest download at http://www.perl.com

The "Unix philosophy" of software construction, at least in the early days of that operating system, was to provide users with a large toolbox of useful "filters"—programs that could do one small task well—and then compose a larger program from the smaller ones. The shell script notations *sh* and *csh* were the means by which composition was done; *sed, awk, tr*, and other programs were some of the more commonly used filters. Perl was developed ostensibly to solve a problem in text processing that *awk* was not good at and has continued to evolve from there.

To summarize Perl completely and succinctly, we probably cannot do much better than this excerpt from the original Unix help file:

> Perl is (an) interpreted language optimized for scanning arbitrary text files, extracting information from those text files, and printing reports based on that information. It's also a good language for many system management tasks. The language is intended to be practical (easy to use, efficient, complete) rather than beautiful (tiny, elegant, minimal). It combines (in the author's opinion, anyway) some of the best features of C, sed, awk, and sh, so people familiar with those

languages should have little difficulty with it. (Language historians will also note some vestiges of csh, Pascal, and even BASIC|PLUS.) Expression syntax corresponds quite closely to C expression syntax. If you have a problem that would ordinarily use sed or awk or sh, but it exceeds their capabilities or must run a little faster, and you don't want to write the silly thing in C, then perl may be for you. There are also translators to turn your sed and awk scripts into perl scripts. OK, enough hype.

Larry Wall is a trained linguist and this interest and expertise shows in Perl. Here he summarizes the nature and intent of his language and his design rationale:

> When they first built the University of California at Irvine campus, they just put the buildings in. They did not put any sidewalks, they just planted grass. The next year, they came back and built the sidewalks where the trails were in the grass. Perl is that kind of a language. It is not designed from first principles. Perl is those sidewalks in the grass. Those trails that were there before were the previous computer languages that Perl has borrowed ideas from. And Perl has unashamedly borrowed ideas from many, many different languages. Those paths can go diagonally. We want shortcuts. Sometimes we want to be able to do the orthogonal thing, so Perl generally allows the orthogonal approach also. But it also allows a certain number of shortcuts, and being able to insert those shortcuts is part of that evolutionary thing.
>
> I don't want to claim that this is the only way to design a computer language, or that everyone is going to actually enjoy a computer language that is designed in this way. Obviously, some people speak other languages. But Perl was an experiment in trying to come up with not a large language—not as large as English—but a medium-sized language, and to try to see if, by adding certain kinds of complexity from natural language, the expressiveness of the language grew faster than the pain of using it. And, by and large, I think that experiment has been successful.
>
> —Larry Wall, in *Dr. Dobbs Journal*, Feb. 1998

In its early versions, Perl was simple and much closer to the scripting notations from which it grew. In later versions, as with many languages, Perl began to accumulate features and facilities as advocates tried to make it more general-purpose and keep it in step with object-oriented language developments.

PERL LANGUAGE OVERVIEW

A discussion of the programming features and facilities of Perl is in order before we present the areas in which Perl can be applied. This will be an overview, not a tutorial, so no attempt is made to provide an exhaustive or in-depth treatment. Components of Perl that are unique or unusual will be emphasized at the expense of features common to many languages.

Perl is an interpreted language, meaning that a control program that understands the semantics of the language and its components (the *interpreter*) executes program components individually as they are encountered in the control flow. Today this usually is done by first translating the source code into an intermediate representation—called *bytecode*—and then interpreting the bytecode. Interpreted execution makes Perl flexible, convenient, and fast for programming, with some penalty paid in execution speed.

Perl programs are often called *scripts* because of its historical development as an extension of the Unix command-level command scripting notations. A Perl script consists of a series of *declarations* and *statements* with interspersed *comments*. A *declaration* gives the interpreter type information and reserves storage for data. Each *statement* is a command that is recognized by the Perl interpreter and executed. Every statement is usually terminated by a semicolon, and in keeping with most modern syntax, white space between words is not significant. A *comment* begins with the # character and can appear anywhere; everything from the # to the end of the line is ignored by the Perl interpreter.

Though the Perl language proper does not have multiline (block) comments, the effect can be achieved using POD (plain old documentation) directives that are ignored by the interpreter. POD directives begin with an "=" and can appear anywhere the interpreter expects a statement to start. They allow various forms of documentation and text markup to be embedded in Perl scripts and are meant to be processed by separate POD tools. A block comment can be made by opening it with a directive such as `"=comment"` and ending it with `"=cut"`. All lines in between are ignored.

Perl is considered by some to have convoluted and confusing syntax; others consider this same syntax to be compact, flexible, and elegant. Though the language has most of the features one would expect in a "full service" programming notation, Perl has become well known for its capabilities in a few areas where it exceeds the capabilities of most other languages. These include

String manipulation
File handling
Regular expressions and pattern matching
Flexible arrays (hashes, or associative arrays)

In the following sections we present the basics of core language, with emphasis on the functions that Perl does especially well.

Basic Data Types and Values

Perl provides three types of data for programmers to manipulate: *scalar, array*, and *hash* (*associative array*). Scalar types are well known to most programmers, as is the array type. The hash is less well known and one of the most powerful aspects of Perl (along with pattern matching, discussed later). Scalar values include *integer, real, string*,

and *boolean*, with the expected operations available for each. Variables do not have to be declared before use; the first character indicates the type.

Scalars

Scalar variables have a leading "$"; for example, these are all valid Perl scalar variables:

```
$n $N $var28 $hello_World $_X_ $_
```

Case matters in Perl, so the first two examples are different variables. The final variable "$_" is special, one of many that the Perl interpreter will use by default in various operations if the programmer does not indicate otherwise. Any particular valid identifier can be used to designate a scalar, an array, *and* a hash. The leading character determines which of the three types the variable has. For example, here the identifier "name" is used to denote three different variables:

```
$name @name %name
```

The first is a scalar; the second is an array; the last is a hash. All three can be used concurrently, as they denote different storage areas. A variable of the type *scalar* can contain any scalar value:

```
$v1 = "good morning";
$v1 = 127;
```

The first assignment places a string value in the variable. The second replaces the string with an integer value. This is different from many strongly typed languages (such as C++ and Java), where types are finely divided into categories such as integer, real, string, and boolean. In Perl these are values, but not types; Perl uses type distinction mainly to separate singular entities (scalar) from collective entities (arrays and hashes).

String values are delimited with either single or double quotes. Single-quoted literals are used exactly as written, whereas double-quoted literals are subject to escape character and variable interpolation before the final value is obtained. For example:

```
$numer = 2;
$st1 = 'one fine $numer day';
$st2 = "$numer fine day \n";
print $st2;
print "$st1\n";
```

The output from this script is

```
2 fine day
one fine $numer day
```

Four interpolations have taken place. In the second line, the `$numer` appears to be a use of a variable, but because the string is in single quotes the characters are included as they appear. In the third line the string literal is in double quotes, so the variable name is replaced with its current value (2) to produce the final string value; the escape sequence "\n" also puts in a *newline* character. The first *print* statement shows the result. The second *print* shows two interpolations as well, as the string being printed is in double quotes. The value in `$st1` is interpolated into the output string and then a *newline* is added to the end. Even though `$st1` has `'$numer'` in its value, this is not recursively interpolated when those characters are put into the output string.

Context

Many of the operators in Perl will work on all three types, with different results being produced depending on the form of the operands. This polymorphism is known in Perl as *context*. There are two major contexts: *scalar* and *list*. Scalar context is further classified as *numeric, string,* or *boolean*. Consider a scalar variable that is assigned an integer value. If the same variable is later used in a string context (meaning operated on by a string function), the integer value is automatically treated as a string of ASCII characters representing the digits of the integer. For example:

```
$v1 = 127;
$v1 = $v1 . ", and more !!";
print $v1, "\n";
$v1 = 127;
print $v1 + " 151 ", "\n";
print $v1 + ", and more !!", "\n";
```

The output from these statements is

```
127, and more !!
278
127
```

The "." operator in the second assignment performs string concatenation, making the expression have string context; the interpreter therefore treats the value of `$v` as an ASCII string of digits, not as an integer. Because integers and strings are both scalars, we can store the resulting string value back into `$v`, which previously held an integer value. The `"+"` in the second *print* is an arithmetic operator, giving that expression numeric context; the interpreter converts the string `"151"` to the obvious integer value for addition. The final *print* is also a numeric context, but there is no obvious valid integer value for the string `," and more !!"` so a zero is used for the addition.

Arrays

Use of array variables in expressions can cause some confusion. In fact, Perl's somewhat convoluted syntax is one of the main complaints against the language. ("It's the magic that counts," quipped Larry Wall on one occasion, when this feature of the syntax was publicly noted.) When an array is manipulated collectively, the leading "@" notation is used:

```
@A = ("hi", "low", 17, 2.14159, "medium");
@A = @B;
print "$B[1] \n";
```

This code fragment outputs "`low`" on one line. The first statement creates an array A by assigning the members of

the list to consecutive array elements; the second line then sets another array B to have all the same values. Finally, it prints the second element from B (array indexes start at 0 in Perl). Arrays contain scalars as elements, so integer, real, boolean, and string values can be stored in the same array. Note also that when individual elements in an array are manipulated, the scalar notation "$" is used; the reasoning is that the element itself is not an array, but a scalar. Using this notation, individual array elements can be given values via assignment.

Array references can be used anywhere a scalar can be, such as in a subscript expression. If an array subscript expression produces a scalar that is not an integer (such as string or real) Perl converts it to some reasonable integer interpretation:

```
$A[0] = 72;
$A[4] = "moderate exercise";
$A[$i] = $B[$j];
print "$A[$A[3]]\n";
```

Here the last line produces "17" on one line. The inner expression evaluates to 2.14159; to use this as a subscript Perl takes the integer part. Thus, it prints `$A[2]`, which is the scalar 17.

Hashes

Hashes (associative arrays) have elements that are indexed by any scalar (usually strings) rather than by integer value:

```
$assoc{"first"} = 37;
$assoc{'second'} = 82;
$assoc{"third"} = "3_rd";
```

Syntactically, subscripts are delimited with curly braces rather than brackets, and string constants used for subscripts can be delimited with single or double quotes. Elements are retrieved from an associative array in similar fashion:

```
$str = "first";
$N = $assoc{$str};
print "$assoc{'second'} - $N \n";
print %assoc, "\n";
```

The output from this segment is

```
82 - 37
first37third3_rdsecond82
```

The last line shows that accessing the entire associative array is possible with the leading "%," and that when printed this way, the output shows up as (index, value) pairs in arbitrary order (determined by the interpreter). Note also in the next to last line that scalar variables can appear inside strings; they are *interpolated*, that is, replaced by their values in the value of the string. Input/output is discussed in more detail later.

Basic Operations

Scalar

Scalar values can be manipulated by the common operators we would expect of a modern programming language. Numbers have arithmetic operations and logical comparisons, autoincrement and decrement (++ and −−), and operator assignments (+ =, − =, ∗ =). Strings have lexical comparison operations, as well as concatenation, truncation, substring extraction and indexing operations. Booleans have the expected logical operators, and the conjunction (&&) and disjunction (||) are evaluated clause by clause and will short-circuit as soon as the final expression value can be determined.

Array

Perl has the expected assignment and referencing operators for arrays; it also provides subrange operators to use part of an array. `$#arr3` will give you the scalar value that is the last index used in array `@arr3`; because Perl indexes arrays from 0, `$#arr3 + 1` will give the array length. Several predefined functions allow a programmer to use arrays as implementations of other data abstractions. For example, *push(@ar7, $elt),* and *pop(@arr7)* will treat the array @arr7 as a stack; *reverse, sort, shift, join, splice,* and *map* are other predefined functions on arrays.

Hash (Associative Array)

Hashes have assignment and multiassignment to create attribute/value pairs and have array referencing via scalar subscripts (usually strings) to retrieve the value associated with an attribute. Most other operations are provided as functions on the array name. For example, *keys (%aa2)* will return a list of the subscript strings for which there are values in the hash *aa2*. Other such operations are *values*, *each*, and *delete*.

Control Flow

Aside from syntactic differences, Perl has much the same *while*, *until*, and *for* loop structures most programming languages have. In keeping with the stated goal of being flexible and not limiting, however, Perl allows several forms of limited jumps within the context of loops that many other languages do not.

If/elsif/else

The traditional if/then/else conditional statement is altered a bit in Perl. There is no then keyword required on the true clause, and following that may be nothing, an else clause, or an elsif clauses. An elsif clause flattens the decision tree that would otherwise be formed by having another if/else as the body of an else clause. Perl lacks a case statement, so the elsif functions in this capacity, as in this example:

```
if ($thresh < 10) {
     # ... the 'then' block of the
        conditional
} elsif ($thresh < 20) {
     # the next block in the decision tree
} elsif ($thresh < 40) {
     # and the next ...
```

```
} else {
     # the final clause catches what falls
        through
}
```

The negation shorthand *unless(exp)* can be used for *if (!exp)* in all contexts where the *if* keyword is valid.

Expressions and Do Blocks

In Perl, statements are viewed as expressions, and executing a statement produces a value for that expression. Every value can also, by convention, be interpreted as a truth value. Any empty string, the number 0, and the string "0" are all treated as "false"; other values are treated as "true" (with a few exceptions). For example, executing the assignment $a = 27$ has the effect of setting the value of variable $a, but it also produces the value 27 as the result of the expression. If this expression were used in a context where a Boolean was needed, then the 27 would be interpreted as "true":

```
$a = $b = 27; # assigns 27 to both
  variables,
              # since the first
                 assignment to $b
                 produces 27 as its value
print "val: ", ($a = $b = 27), "\n";
if ($a = 27) {# assignment to $a...
   illustration only, not good style
    print "it was true \n";
} else {
    print "it was false \n";
}
if ($a = 0) {# another assignment to $a
    print "it was true \n";
} else {
    print "it was false \n";
}
```

This code fragment produces this output:

```
val: 27
It was true
It was false
```

A *do* {*BLOCK*} statement simply executes the code within the statement block and returns the value of the last expression in the block. We can use this feature combined with statement values to produce an alternate form of conditional. The following two statements are equivalent:

```
($thresh < 125) && do {print "it passed
  \n";};
if ($thresh < 125) {print "it passed \n";};
```

In the first form we also make use of the fact that Perl will evaluate the clauses of a logical conjunction one at a time, left to right, and stop if one should evaluate to false. In this case, should the boolean comparison fail, the second clause of the conjunction (the one with the printing) will not be attempted.

Loop Structures

Looping in Perl is done with variants of the *while*, the *do,* and the *for* structures. The *while* structure is equivalent to that of Java, C, or C++. The loop body block executes as long as the controlling expression remains true. The *until(expnB)* structure is functionally equivalent to *while(! expnB)*:

```
while ($d < 37) {$d++; $sum += $d;}
until ($d >= 37) {$d++; $sum += $d;}
```

The *do/while* and *do/until* structures work similarly to the *while* structure, except that the code is executed at least once before the condition is checked.

```
do {$d++; $sum += $d;} while ($d < 37);
do {$d++; $sum += $d;} until ($d >= 37);
```

The *for* structure works similarly to those of C, C++, or Java and is really syntactic sugar for a specific type of *while* statement. More interesting is the *foreach* loop, which is specifically designed for systematic processing of Perl's native data types. The *foreach* structure takes a scalar, a list, and a block and executes the block of code, setting the scalar to each value in the list, one at a time. Thus the *foreach* loop is a form of *iterator*, giving access to every element of some controlling collection. Consider this example:

```
my @collection = ("first", "second",
   "third", "fourth");
foreach $item (@collection) {print
   "$item\n";}
```

This will print out each item in collection on a line by itself. We are permitted to declare the scalar variable directly within the *foreach*, and its scope is the extent of the loop. Perl programmers find the *foreach* loop to be one of the most useful structures in the language.

last Operator

The *last* operator, as well as the *next* operator that follows, applies only to loop control structures. These operators cause execution to jump from where they occur to some other position, defined with respect to the block structure of the encompassing control structure. Thus, they function as limited forms of *goto*. Last causes control to jump from where it occurs to the first statement following the enclosing block. For example:

```
$d = 2;
while ($d++) {
    if ($d >= 37) { last;}
    $sum += $d;
}
# last jumps to here
```

next operator

The *next* operator is similar to *last* except that execution jumps to the end of the block, but remains *inside* the block,

rather than exiting it. Thus, iteration continues normally. For example:

```
while ($d < 37) {
   $d++;
   if (($d%5)==1 ) { next};
   $sum += $d;
   # next jumps to here
}
```

Jumps can be made from inner nested loops to points in outer loops by labeling the loops, and using the appropriate label after the *last* and *next*. We can now combine several of these features to give another way to "fake" the case statement shown previously as a decision tree with *if / elseif /else*:

```
CASE: {
   ($thresh < 10) && do {
      #  the 'then' block of the conditional
      last CASE; }
   ($thresh < 20) && do {
      #  the next block in the
         decision tree
      last CASE; }
   ($thresh < 40) && do {
      #  and the next...
      last CASE; }
   #  the final clause here catches what
     falls through
} # end of CASE block
```

As we mentioned earlier, there is *always* more than one way to do things in Perl.

Subroutines

A subprogram in Perl is often called a function, but we shall use the term subroutine here to distinguish programmer-defined structures from the built-in functions of Perl. A subroutine is invoked within the context of some expression. In early versions of Perl, an ampersand (&) was placed before the subroutine name to denote invocation; current versions allow invocation without it as well. If the subroutine takes arguments, they are placed within parentheses following the name of the subroutine.

```
&aSubProg();
bSubProg();
cSubProg($ar3, $temp5, @ARY);
```

Control is transferred to the code of the subroutine definition, and transfers back either when the end of the subroutine code is reached, or an explicit return()statement is executed in the subroutine body.

The subroutine *definition* is marked by the keyword `sub` followed by the name of the subroutine, without an ampersand prefix. A block of code for the subroutinebody follows, enclosed in curly braces; this is executed when the subroutine is called.

```
sub aSubProg {
   stmt_1;
   stmt_2;
   $a = $b + $c;
}
```

The value returned by a Perl subroutine is the value of the last expression evaluated in the subroutine. In this example, *aSubProg* will return the value `$a` has at the time when the subroutine ends. Functions such as *print* return values of `0` or `1`, indicating failure or success.

Arguments are enclosed in parentheses following the name of the subroutine during invocation; thus, they constitute a *list*. They are available within the subroutine definition block through `@_` the predefined (list) variable:

```
aSubProg ($a, "Literal_string", $b);

sub aSubProg {
   foreach $temp(@_) { print "$temp \n"; }
}
```

Any variables defined within the body of a Perl program are available inside a Perl subroutine as global variables. Consequently, Perl provides an explicit scope operator (*my*) that can be used to limit the visibility of variables and protect globals from inadvertent side effects. Similarly, these locals will not be visible outside the subroutine. Local variables are, by convention, defined at the top of a Perl subroutine:

```
aFunc ($a, $b);
sub aFunc {
   my ($aLocal, $bLocal);
   $aLocal = $_[0]; # @_ is used $_[i] for
     individual arguments
   $bLocal = $_[1];
}
```

`$aLocal` and `$bLocal` will have the same values inside the subroutine as `$a` and `$b` have at the time it is invoked. Changes to either local variable inside the function, however, will not affect the values of `$a` or `$b`.

Built-In Functions and System Operations

Perl offers a rich selection of built-in functions as part of the standard interpreter. These include mathematical operations (such as abs, sin, sqrt, log); list manipulation operations (such as join, reverse, sort); array manipulation operations (such as push, pop, shift); string manipulation operations (such as chop, index, length, substr, pack, reverse); and myriad operating system functions reflecting Perl's Unix birthright.

Because one of the reasons for the creation of Perl was to give Unix programmers more expressive power and convenience, the language provides several mechanisms for invocating operating system services from executing

scripts. The most general method is the system function:

```
$retVal = system("pwd");
```

In this example, the Perl interpreter uses the system command to get the underlying operating system to execute the Unix "pwd" command. The result of the command appears on STDOUT just as it would if it were done from the command line; the return value, in this case, is an indicator of success or failure. Often programmers want to capture the output of a system command for inclusion in the executing script. This is accomplished by enclosing the command in backward single quotes, often called "backticks":

```
$dir = 'pwd';
print "the current directory is $dir \n";
```

Many other operating system (specifically, Unix) manipulations are available in Perl via built-in functions. The *chdir* function allows a Perl script to alter the default directory in which it finds its files while executing; the *opendir, readdir*, and *closedir* functions allow a Perl script to obtain directory listings; *mkdir* and *rmdir* allow a script to create and delete directories; *rename* and *chmod* allow a script to rename a file and change its access permissions. All these capabilities exist because Perl was originally designed to make it easy for system managers to write programs to manipulate the operating system and user file spaces.

Functions *exec, fork, wait*, and *exit* allow scripts to create and manage child processes. Perl provides a means of connecting a running process with a file handle, allowing information to be sent to the process as input using print statements, or allowing the process to generate information to be read as if it were coming from a file. We illustrate these features in the section *Network Programming in Perl*.

Regular Expressions and Pattern Matching

Perhaps the most useful, powerful, and recognizably Perlish aspect of Perl is its pattern-matching facilities and the rich and succinct text manipulations they make possible. Given a pattern and a string in which to search for that pattern, several operators in Perl will determine whether—and if so, where—the pattern occurs. The pattern descriptions themselves are called *regular expressions*. In addition to providing a general mechanism for evaluating regular expressions, Perl provides several operators that perform various manipulations on strings based upon the results of a pattern match.

Regular Expression Syntax

Patterns in Perl are expressed as regular expressions, and they come to the language through its Unix *awk* heritage. Because regular expressions are well understood from many areas of computing, we will not give an involved introduction to them here. Rather, we will simply use Perl examples to give an idea of the text processing power they give the language.

By default, regular expressions are strings that are delimited by slashes, *e.g.*, `/rooster/`. This delimiter can be changed, but we will use it for the examples. By default, the string that will be searched is in the variable `$_`. One can apply the expression to other strings and string variables, as will be explained below.

The simplest form of pattern is a *literal string*. For example:

```
if (/chicken/) {print "chicken found in
   $_\n";}
```

The "/" delimiters appearing alone denote a default application of the match operator. Thus this code fragment searches in the default variable $_ for a match to the literal "chicken," returning true if found. In addition to including literal characters, expressions can contain categories of characters. They can specify specific sequences with arbitrary intervening strings; they can specify matches at the beginning or end; they can specify exact matches, or matches that ignore character case. Examples of these uses include:

```
/.at/              # matches "cat," "bat," but not "at"
/[aeiou]/          # matches a single character from the set of vowels
/[0-9]/            # matches any single numeric digit
/\d/               # digits, a shorthand for the previous pattern
/[0-9a-zA-Z]*/     # matches a string of alphanumeric characters, or length zero or more
/\w/               # words, a shorthand for the previous pattern
/[^0-9]/           # not a digit
/c*mp/             # any number of c's followed by mp
/a+t/              # one or more a's followed by t
/a?t/              # zero or one a followed by t
/a{2,4}t/          # between 2 and 4 a's followed by t
/k{43}/            # exactly 43 occurrence of "k"
/(pi)+(sq)*/       # strings with one or more "pi" pairs followed by zero or more "sq" pairs
/^on/              # match at start: "on the corner" but not "Meet Jon"
/on$/              # match at end: "Meet Jon" but not "on the corner"
/cat/i             # ignore case, matches "cat", "CAT", "Cat", etc.
$A =~/pong/           # does the content of string variable $A contain "pong"?
<STDIN> =~/b.r+/      #  does the next line of input contain this pattern
                      # which matches bar, bnr, bor, brrr, burrrrrr, etc.
```

Pattern matching is *greedy*, meaning that if a pattern can be found at more than one place in the string, the leftmost instance is returned; if there are overlapping leftmost instances, the longest match will be identified.

String Manipulation

Regular expression operators include a regular expression as an argument but instead of just looking for the pattern and returning a truth value, as in the examples above, they perform some action on the string, such as replacing the matched portion with a specified substring (like the well-known "find and replace" commands in word processing programs). The simplest is the "m" operator, the explicit match. In the following example, a string is searched for the substring "now" (ignoring character case); the match operator return value is interpreted as a Boolean for control of the conditional:

```
my($text) = "Now is the time, now seize
   the day";
if ($text =~ m/now/i) {print "yep, got
   it\n";}
if ($text =~ /now/i) {print "yep, got
  it\n";} # equivalent form, no "m"
```

In general, in invoking the match operator the "m" is usually omitted, as illustrated in the third line above. If a pattern is given with no explicit leading operator, the match operator is employed by default. Though we do not extract or use the matching substring in this example, the operator actually matches on the first three characters "Now" because of the ignore case option.

The substitution operator "`s`" looks for the specified pattern and replaces it with the specified string. By default, it does this for only the first occurrence found in the string. Appending a "`g`" to the end of the expression causes global replacement of all occurrences.

```
s/cat/dog/   # replaces first "cat" with
   "dog" in the default variable $_
s/cat/dog/gi # same thing, but applies
   to "CAT", "Cat" everywhere in $_
$A =~ s/cat/dog/ # substitution on the
   string in $A rather than the default $_
```

The *split* function searches for all occurrences of a pattern in a specified string and returns the pieces that were separated by the pattern occurrences as a list. If no string is specified, the operator is applied to `$_`.

```
$aStr = "All category";
@a = split(/cat/, $aStr); # a[1] is "All "
   and a[2] is "egory"
@a = split(/cat/);        # this split
   happens on the string in default $_
```

The *join* function performs the opposite of a split, assembling the strings of a list into a single string with a separator (the first argument) placed between each part:

```
$a = join(":", "cat", "bird", "dog");
  # returns "cat:bird:dog"
$a = join("", "con", "catenate");
   # returns "concatentate"
$a = "con". "catenate"; # $a gets the value
   "concatentate"
@ar = ("now", "is", "the", "time");
$a = join "", @ar;          # $a gets the
   value "nowisthetime"
```

In the second line above, where the separator is no character at all, the effect of the *join* is the same as using Perl's concatentation operator, as shown in the third line. The added power of *join* is that it will operate on all elements of a list without them being explicitly enumerated, as illustrated in the fourth and fifth lines.

Pattern Memory

The portion of the string that matches a pattern can be assigned to a variable for use later in the statement or in subsequent statements. This feature is triggered by placing the portions of a pattern to be *remembered* in parentheses. When used in the same statement or pattern, the matched segment will be available in the variables `\1, \2, \3`, etc. in the order their targets occur. Beyond the scope of the statement, these stored segments are available in the variables `$1, $2, $3,` etc. as well as contextually. Other matching information available in variables include `$&`, the sequence that matched; `$'`, everything in the string up to the match; and `$'`, everything in the string beyond the match.

For example, the following program separates the file name from the directory path in a Unix-style path name. It works by exploiting Perl's greedy matching, along with the pattern memories:

```
my($text) = "/tmp/subsysA/user5/fyle-zzz";
my($directory, $filename) = $text =~ m/
  (.*\/)(.*)$/;
print "D=$directory, F=$filename\n";
```

The pattern finds the last occurrence of "/" in the target string so that the Unix directory can be split out from the file name. The first set of parentheses saves this directory substring, and the second set captures the file name. The assignment after the match on `$text` stores both pattern memories by positional order into the variables `$directory` and `$filename`. Here is another example using the `\1` and `$1` memory notations:

```
$A = "crave cravats";
$A =~ s/c(.*)v(a.)*s/b\1\2e/;
   # \1 is "rave cra" and \2 is "at"
print "$A\n";
print "$1\n";
print "$2\n";
```

The output from this code fragment is

```
brave craate
rave cra
at
```

The substitute operator in the second line performs the match by first finding the longest string of characters between the "c" and "v" and saving it in the `\1` memory. It then finds the longest string of "a" followed by any single character in the rest, and saves that in the `\2` memory. Once matched, the replacement string is formed by concatenating the memories, adding a "b" at the front, and adding an "e" at the end. The last two lines show that the string parts that matched the pattern parts are still available after the match for as long as the variables `$1` and `$2` are not overwritten.

Input/Output and File Handling

File handling is another area where Perl makes life easy for the programmer. The basic file manipulation operators, coupled with array capabilities, make creating internal structures out of text input succinct and efficient. Files are accessed within a Perl program through *filehandles*, which are bound to a specific file through an `open` statement. By convention, Perl filehandle names are written in all uppercase, to differentiate them from keywords and function names. For example:

```
open (INPUT, "index.html");
```

associates the file named "index.html" with the filehandle INPUT. In this case, the file is opened for read access. It may also be opened for write access and for update (appending) by preceding the filename with appropriate symbols:

```
open (INPUT, ">index.html");  # opens for
   write
open (INPUT, ">>index.html"); # opens for
   appending
```

Because Perl will continue operating regardless of whether a file open is successful or not, we need to test the success of an `open` statement. Like other Perl constructs, the `open` statement returns a true or false value. Thus, one common way to test the success of the open and take appropriate action is to combine the lazy evaluation of logical *or* with a `die` clause, which prints a message to STDERR and terminates execution:

```
open (INPUT, "index.html") || die "Error
   opening file index.html ";
```

Files are closed implicitly when a script ends, but they also may be closed explicitly:

```
close (INPUT);
```

Perl provides default access to the keyboard, terminal screen, and error log with predefined filehandles STDIN, STDOUT, and STDERR; these handles will be automatically available whenever a script is executed. Once opened and associated with a filehandle, a file can be read with the diamond operator (<>), which can appear in a variety of constructs. STDIN is most commonly accessed this way. When placed in a scalar context, the diamond operator returns the next line; when place in an array context, it returns the entire file, one line per item in the array. For example:

```
$a = <STDIN>; # returns next line in file
@a = <STDIN>; # returns entire file
```

STDOUT is the default file accessed through a `print` statement. STDERR is the file used by the system to which it writes error messages; it is usually mapped to the terminal display. Here is an example that reads an entire file from STDIN, line-by-line, and echos each line to STDOUT with line numbering:

```
$lnum = 0;
while (<STDIN>) {  # read one line at a time
  until EOF
                  # in this case, the
                    default variable $_
                    receives the line
   chomp;         # remove line-ending
     character (newline here)
                  # again, it operates on $_
                     automatically
   $lnum++;       # auto-increment operator
     on line counter
   print "$lnum: $_\n";  # print the line
     read, using default $_
}
```

This shows one of the many Perl conveniences. In many contexts, if no scalar variable is indicated, an operation will give a value to a variable named '`$_`', the default scalar variable. This is in keeping with Perl's design philosophy of making it very easy to do common tasks. We could also have omitted the filehandle STDIN and simply have written "`while (<>)`"; the diamond operator will operate on STDIN by default if given no filehandle explicitly.

Once a file has been opened for either write or update access, data can be sent to that file through the *print* function. A *print* with no filehandle operates on STDOUT by default. For example:

```
print OUTPUT "$next \n"; # to a file opened
   with handle OUTPUT
print "this statement works on STDOUT
   by default\n";
```

In many circumstances the actions taken by a Perl program should take into account attributes of a file, such as whether or not it currently exists, or whether it has content. A number of tests can be performed on files through *file test* operators. For example, to check for file existence use the *–e* test:

```
if (-e "someFile.txt") {
   open (AFYLE, "someFile.txt") || die "not
     able to open file";
}
```

Using different characters, many other attributes can be tested including if a file is readable, writable, executable, or owned by certain users, if it is text or binary, if it is a directory or symbolic link, or if it is empty, to name a few.

There's More Than One Way To Do It

In concluding this section we again illustrate the famous Perl adage, this time with file *open* statements. Here are several examples of conditional expressions for safely opening files and trapping errors.

```
$aFile = "foo.txt";
if (!open(fh, $aFile)) {die "(a) Can't open
   $aFile: $!";}
open(fh,$aFile) ? " : die "(d) Can't open
   $aFile: $!";
die "(b) Can't open $aFile:
   $!" unless open(fh,$aFile);
open(fh,$aFile) || die "(c) Can't open
   $aFile: $!";
```

The last four lines all do the same thing.

Other Perl Features

Perl has several other features and capabilities that have found their way into the language as it evolved. These later features tend to be capabilities that programmers found useful in other languages and desired to have in Perl. In particular, Perl version 5 introduced *classes, objects,* and *references* (or pointers) into a language that was previously a more traditional Unix scripting notation "on steroids." Because they do not greatly enhance Perl's capabilities in the areas for which it has proven especially superior (text processing, file handling, string matching, OS interactions) we will not go into them in detail. Some programmers even consider these additions to aggravate the already difficult task of reading Perl code. These features are not *unimportant* aspects of the language; they are simply well beyond the original domains of expertise and applicability for which Perl was developed. As such, they represent the natural end to which languages tend to evolve as they gain popularity—something of everything for everyone.

Perl has many more sophisticated capabilities. Access to the interpreter is available to an executing script through the *eval* function, allowing a program to create and then run new code dynamically. Symbol tables can be accessed and manipulated directly with Perl *typeglobs*. Function *closures* can be created (as in many functional languages) allowing subroutines to be packaged dynamically with their data and passed back from a function call as a reference for execution later. *Packages* and *modules* provide encapsulation and namespace control. The later versions of Perl even support concurrent computations with a *thread* model.

We refer the reader to the texts cited in *For More Information* for thorough presentations of all these topics.

PUTTING IT ALL TOGETHER: SAMPLE PROGRAMS

First Example: Text Processing

Here is a Perl script that will take as input a file called "foo.txt" and produce as output a file called "bar.txt"; lines in input will be copied to output, except for the following transformations:

any line with the string "IgNore" in it will *not* go to output
any line with the string "#" in it will have that character and all characters after it, to end of line, removed
any string "*DATE*" will be replaced by the current date in output

One program to do this is as follows:

```
#!/usr/local/bin/perl
$infile = "foo.txt";
$outfile = "bar.txt";
$scrapfile = "baz.txt";
open(INF",<$infile") || die "Can't open
   $infile for reading";
open(OUTF",>$outfile") || die "Can't open
   $outfile for writing";
open(SCRAPS",>$scrapfile") || die "Can't
   open $scrapfile for writing";
chop($date = 'date'); # run system command,
   remove the newline at the end
foreach $line (<INF>) {
   if ($line =~ /IgNore/) {
       print SCRAPS $line;
       next;
   }
   $line =~ s/\*DATE\*/$date/g;
   if ($line =~ /\#/) {
       @parts = split ("#", $line);
       print OUTF "$parts[0]\n";
       print SCRAPS "#". @parts[1..$#parts];
         # range of elements
   } else {
       print OUTF $line;
   }
}
close INF; close OUTF; close SCRAPS;
```

In keeping with the Perl adage that there's more than one way to do things, here is an alternative way to write the *foreach* loop; this one uses the implicit `$_` variable for pattern matching:

```
# this version uses the implicitly defines
  $_ variable
foreach (<INF>) {
   if ( /IgNore/ ) {
       print SCRAPS;
       next;
       }
   s/\*DATE\*/$date/g;
   if ( /\#/ ) {
       @parts = split ("#");
```

```
      print OUTF "$parts[0]\n";
      print SCRAPS "#". @parts[1..$#parts];
        # range of elements
   } else {
      print OUTF;
      }
}
```

And finally, a third version, using boolean and conditional expressions in place of *if–else* statements:

```
# this version uses boolean interpretation
  of expressions as
# substitution for  if clauses in previous
  versions
foreach (<INF>) {
   /IgNore/ && do {print SCRAPS; next};
   s/\*DATE\*/$date/g;
   /#/ ? do {
       @parts = split ("#");
       print OUTF "$parts[0]\n";
       print SCRAPS "#". @parts[1..$#parts];
         # range of elements
   }
   : do {
       print OUTF;
   }
}
```

A Simpler, More Sophisticated Example

Consider this problem: take an input file and produce an output file which is a copy of the input with any duplicate input lines removed. Here is a first solution:

```
#!/usr/local/bin/perl
foreach (<STDIN>) {print unless $seen
  {$_}++;}
```

This is, of course, exactly why so many like Perl so fervently. A task that would take many lines of C code can be done in Perl with a few lines, thanks to the sophisticated text handling facilities built into the language. In this solution, we are reading and writing standard input and output; in Unix we supply specific file names for these streams when the program it is invoked from the command line, like this:

```
second.pl <foo.txt >bar.txt
```

Here is a second solution:

```
#!/usr/local/bin/perl
# this version prints out the unique lines
   in a file, but the order
# is not guaranteed to be the same as they
   appear in the file
foreach (<>) {$unique{$_} = 1;}
print keys(%unique); # values(%unique)
   is the other half
```

And a third solution:

```
#!/usr/local/bin/perl
# this version eliminates duplicate lines
# and prints them out in arbitrary order
# also tells how many time each line was
   seen
# oh, and it sorts the lines in alpha order
foreach (<>) {$unique{$_} += 1;}
foreach (sort keys(%unique)) {
  print "($unique{$_}):$_";
}
```

This last example shows the considerable power and terseness of Perl. In essentially four lines of code, we filter a file to remove duplicate lines, report a count of how many times each unique line appeared in the original input, and print the unique lines sorted in alphabetic order. All the facilities used in this program are part of the standard Perl language definition. It does not depend on any user-supplied routines or libraries.

Directory Information Processing

This example shows more complicated use of pattern memories in text processing. The script reads standard input, which will be piped test from a Linux *dir* command (directory). It writes to standard out, and produces an executable script (in *csh* notation) that copies every file older than 11/01/93 to a directory called `\ancient`. The input looks like this:

```
.            <DIR>        12-18-97 11:14a  .
..           <DIR>        12-18-97 11:14a  ..
INDEX    HTM        3,214  02-06-98  3:12p  index.htm
CONTACT  HTM        7,658  12-24-97  5:13p  contact.htm
PIX          <DIR>        12-18-97 11:14a  pix
FIG12    GIF          898  06-02-97  3:14p  fig12.gif
README   TXT        2,113  12-24-97  5:13p  readme.txt
ACCESS   LOG       12,715  12-24-97  5:24p  ACCESS.LOG
ORDER    EXE       77,339  12-24-97  5:13p  order.exe
         6 file(s)       103,937 bytes
         3 dir(s)     42,378,420 bytes free
```

The Perl solution uses regular expressions, pattern matching, and pattern memories:

```
my $totByte = 0;
while(<>){
   my($line) = $_;
   chomp($line);
   if($line !~ /<DIR>/) { # we don't want to
     process directory lines
       # dates is column 28 and the filename
         is column 44
       if ($line =~ /.{28}(\d\d)-(\d\d)
         -(\d\d).{8}(.+)$/) {
            my($filename) = $4;
            my($yymmdd) = "$3$1$2";
            if($yymmdd lt "931101") {
                print "copy $filename \ \
                  ancient\n";}}
       if ($line =~ /.{12}((\d ||,)
         {14}) \d\d-\d\d-\d\d/) {
```

```
            my($bytecount) = $1;
            $bytecount =~ s/,//; # delete
              commas
            $totByte += $bytecount;
        }
    }
    print STDERR "$totByte bytes are in this
      directory.\n";
}
```

In the first match, the variables `$1`, `$2`, `$3`, and `$4` are the pattern memories corresponding to the parenthesis sets. The first three are reassembled into a *yymmdd* date string which can be compared with the constant "971222." The fourth holds the filename that will be copied to the `\ancient` directory. As a side effect of processing the directory listing, we set up an accumulator and extract a cumulative byte count. This is done with a second match on the same input line, as well as a substitution operation to remove commas from the numbers found.

NETWORK PROGRAMMING IN PERL

The World Wide Web is the most widely known Internet application. Many Web sites provide more than static HTML pages. Instead, they collect and process data or provide some sort of computational service to browsers. For example, several companies operate Web sites that allow a user to enter personal income and expense information, and will then not only compute income tax returns online but also electronically file them with the IRS. There are numerous technical ways to provide this processing horsepower to a Web site (e.g., Microsoft's Active Server Pages, JavaScript, C/C++/C# programs, etc.) but Perl is the most widespread and popular of these options. In this section we look at how Perl scripts can provide communications between Web browsers and servers, and how they can make use of databases for persistent storage. We also discuss some of Perl's capabilities for interprocess communication and network computations.

Web and network programming is not usually done from scratch, but rather by reusing excellent Perl modules written by other programmers to encapsulate details and provide abstractions of the various domain entities and services. We begin with a discussion of the *Comprehensive Perl Archive Network (CPAN)*, the Perl community's repository for these freely shared modules.

Perl Modules and CPAN

CPAN is a large collection of Perl code and documentation that has been donated by developers to the greater Perl programming community. It is accessed on the Web at http://www.cpan.org and contains many modules that have become *de facto* standards for common Perl scripting tasks. In addition to programmer-contributed modules, the source code for the standard Perl distribution can be found there. In the words of Larry Wall in *Programming Perl*, "If it's written in Perl, and it's helpful and free, it's probably on CPAN."

Modules

The *module* is the main mechanism for code reuse in Perl. A module is a package (protected namespace) declared in a file that has "`.pm`" as its filename extension; the package, module, and file have the same name. The author of a module defines which names within it are to be made available to outside Perl programs. This is done through the *Export* module. To incorporate the variables, subroutines, and objects of a module into a program, the *use* statement is employed:

```
use JacksCode; # in which a variable
   $jackrabbit is declared
print "$jackrabbit \n";
print "$JacksCode::jackrabbit \n";
```

In this example, the *use* statement requests access to all names that are exported from the module "JacksCode," which the interpreter will expect to find in a file named "JacksCode.pm" someplace on its search path. If this module declares a variable named "`$jackrabbit`" then the last two lines do the same thing. A variable name imported from a module need no longer be fully qualified with the module name. There are several alternate forms of the *use* statement that give finer-grained control over which of the exported names are imported.

Many of the modules most commonly used by programmers come as part of the standard Perl distribution. CPAN contains dozens of others, including:

CGI, HTML, HTTP, LWP, Apache module families for Web server scripts

POSIX for Unix-programming compatibility

Net::FTP, Net::DNS, Net::TCP, Net::SMTP, Net::IMAP, and many other for dozens of protocols

Math::BigInt, Math::Trig, Math::Polynomial, Statistics and dozens more supporting various forms of mathematical structures and functions

List, Set, Heap, Graph module families giving common abstract data types

Date, Time, Calendar module families

Language::ML, Language::Prolog, C::DynaLib, Python, Java, and other language interfaces

PostScript, Font, PDF, XML, RTF, Tex, SQL module families for documents

PGP, DES, Crypt, Authen module families for encryption and security

As enjoyable as Perl programmers find their craft to be, no one wants to spend time rewriting code someone else has already done well. CPAN is the result of an enthusiastic community effort to leverage success.

Web Server Scripts with CGI

When a Web page contains fill-out forms, or has some other computational behavior required, there are several ways to provide the processing needed on the Web server side of the transaction. One way is via scripts that adhere to the data formatting standards of the CGI Web interface. CGI scripts can be written in any programming language that the server will support. A separate chapter

in this compilation covers the CGI standard in detail, so we concentrate here on the specifics of how Perl allows programmers to take advantage of this standard for Web development.

In any interaction between a Web browser and a Web server, there are data being exchanged. The browser sends information to the server requesting that some action be taken, and the server sends a reply back, usually in the form of an HTML Web page. This interaction often happens by the user filling out the fields of a form in the browser window and clicking on some "submit" button. Submitting the form entails the browser collecting the data from the form fields, encoding it as a string according to the CGI standard, and passing it to the Web server specified in the URL associated with the submit button in the form. This URL contains not only the location of the server that will process the form data, but also the name of the Perl script that should be executed as well on the server side.

Module "CGI"

The data from the browser form is made available to Perl via *environment variables*. In the early days of the Web, site programmers would "roll their own," writing collections of Perl subroutines to decode the incoming CGI-compliant data and process them in various ways. Today, the task is made much easier through the Perl module "CGI," which provides a *de facto* programming standard for these server-side functions. Using the CGI module, a form processing script looks something like this (simplified):

```
use CGI;
$q = CGI::new();
$mid = $q->param("measid");
$uid = $q->param("uid");
$pwd = $q->param("passwd");
print $q->header();
print $q->head($q->title("Company
  Evaluation"));
print $q->body(
   $q->h1("$uid: Submit My Report"),
   $q->hr,
      etc... rest of body elements...
   );
)
```

(The arrow notation (->) is the Perl syntax for dereferencing a reference (chasing a pointer). In this module, and others following, it is used to access the fields and functions of a Perl object.)

As shown here, the CGI module provides functions for retrieving environment variables, creating Web forms as output, and generating HTML tags. This example is selecting data from a Web form containing text fields called "measid," "uid," and "passwd." It is generating an HTTP-compliant return message with the necessary header and an HTML page for the browser to display. Assuming the "uid" here comes in from the form as "Jones," we get:

```
Content-type: text/html;
  charset=ISO-8859--1
<head>
   <title>Company Evaluation</title>
</head>
<body>
   <h1>Jones: Submit My Report</h1>
   <hr>
      etc...
```

The CGI module also assists the Web site developer in solving other problems common with Web scripting. Because the HTTP protocol is stateless, one problem is maintaining session state from one invocation of a script to the next. This is normally done with *cookies,* data a server asks the Web browser to store locally and return on request. However, not all browsers allow cookies, and in those that do the user may turn cookies off for security or privacy reasons. To help with this a script using CGI, when called multiple times, will receive default values for its input fields from the previous invocation of the script.

Web Clients with LWP

Whereas the CGI module supports construction of scripts on the server side of a Web connection, the modules in *LWP (Library for Web access in Perl)* provides support for developing applications on the client side. Most notable among Web clients are the GUI-based browsers, but many other applications acts as clients in HTTP interactions with Web servers. For example, Web crawlers and spiders are non-GUI programs (called "bots" or robots) that continuously search the Web for pages meeting various criteria for cataloging.

The different modules in LWP support different aspects of Web client construction and operation:

HTML for parsing and converting HTML files
HTTP for implementing the requests and responses of the HTTP protocol
LWP core module, for network connections and client/server transactions
URI for parsing and handling URLs
WWW implementing robot program standards
Font for handling Adobe fonts
File for parsing directory listings and related information

A Web interaction starts with a client program establishing a network connection to some server. At the low level this is done via sockets with the TCP/IP protocol. Perl does support socket programming directly (see below), and the module *Net* contains functions to allow a program to follow TCP/IP (as well as many others Internet protocols, such as FTP, DNS, and SMTP). On top of sockets and TCP/IP for basic data transfer, the HTTP protocol dictates the structure and content of messages exchanged between client and server.

Rather than deal with all these technologies individually, the *LWP::UserAgent* module allows the programmer to manage all client-side activities through a single interface. A simple client script would look like this:

```
use LWP::UserAgent; # imports other
  modules too
$client = new LWP::UserAgent();
$acmeReps = new URI::URL('www.acme.com/
  reports/index.html');
$outHead =
   new HTTP::Headers(User-Agent=>'MyBot
     v2.0', Accept=>'text/html');
$outMsg = new HTTP::Request(GET, $acmeReps,
  $outHead);
$inMSg = $client->request($outMsg);
$inMsg->is_success ? {print $inMsg->
  content;} : {print $inMsg->message;}
```

The network connections and message formatting to HTTP protocol requirements is handled transparently by the LWP functions. Here, the client causes a request like this to be sent to the Web server at www.acme.com:

```
GET/reports/index.html HTTP/1.0
User-Agent: MyBot v2.0
Accept: text/html
```

If the requested file is not there, the response from the server will be an error message. If it is there, the response will contain the HTML file for the "bot" to process in some fashion.

Database Use

Many Web sites depend on databases to maintain persistent information—customer data, statistics on site usage, collection of experimental results, medical records. Perl has several ways a CGI script (or any Perl script, Web-based or otherwise) can store and retrieve database records.

Module "DBM"

Perl comes with a module called DBM (database module) that contains functions implementing a built-in database structure. DBM treats an external data store internally as hashes, or key/value pairs. This internal database is intended for simple, fast usage; searching requires as few as three lines of script. The method is convenient for fast retrieval, even from very large databases, when there is a unique key for searching (e.g., ISBN numbers). It is not as useful for complex data or queries:

```
dbmopen (%db, $database, 0400) || die "Can't
  open DB";     #open read only
for ($k=$max; $k< $max+20; $k++){print
  "$k $db{$k}"} #get and print data
dbmclose (%db); #close database
```

In this example, we know the index values are numbers and are unique. The database looks to the Perl script like a hash, so the associative array variable `%db` is used to get the data values from it.

One useful feature Perl provides is DBM filters. These are small data transformation routines, written by a programmer to manage situations where the format of the data fields in a database is not quite compatible with the form needed by a script. Rather than put small data manipulation code chunks scattered throughout the script, or write and call extra functions, DBM filters can be attached directly to the fields of a database; data transformation then takes place automatically as field values are moved into and out of the database. This feature makes the code using the database easier to read and less error prone due to less code replication.

Module "DBI"

For more advanced applications, a relational database is often more desirable than the simple structure of DBM. For this Perl provides the DBI module, or data base interface. DBI is designed to hide the details of specific database systems, allowing the programmer to create scripts that are general. The interface allows expressing SQL queries, executing them against a specific database, and retrieving the results. The DBI module does not contain code specific to any database vendor's product, but it does have references to numerous vendor-specific modules called DBD's (database drivers). A DBD module will contain the detailed code needed to communicate with a specific brand of database system.

Figure 1 illustrates the relationship among the executing script, the DBI, the DBD, and the physical database. When a Perl script invokes a DBI function to execute a query, the query is routed to the appropriate DBD module according to how the DBI database handle was opened (as an Oracle DB, as an Access DB, etc.). The DBD module communicates with the actual files or tables of the physical database system and produces query results. These results are communicated back to the DBI module, which relays them back to the user's Perl script. This layer of indirection gives Perl scripts a generality that makes migration from one physical DB system to another relative painless.

Module "ODBC"

In addition to the DBI module, programmers wishing to write Perl scripts that interface to external databases, such as Access or Oracle, can obtain an ODBC compatibility package as a free download from several third party distributors. This module contains functions that implement the ODBC standard database interface. Perl scripts written to use this standard can then work with any relational database under them, as long as that database has its own ODBC interface. Almost all major relational database vendors provide an ODBC implementation for their products. ODBC provides the same advantages as DBI, but the ODBC standard was designed outside Perl and is available in many other programming languages as well.

Processes and IPC

Although the Web is the major Internet application, it is certainly not the only one, and Perl includes facilities to support general network applications. Although not designed specifically for concurrent or multiprocess computations, Perl does support various forms of processes, interprocess communication (IPC), and concurrency. (Perl having been born of Unix, this section is heavy on Unix process concepts such as pipes and forking.) Processes are

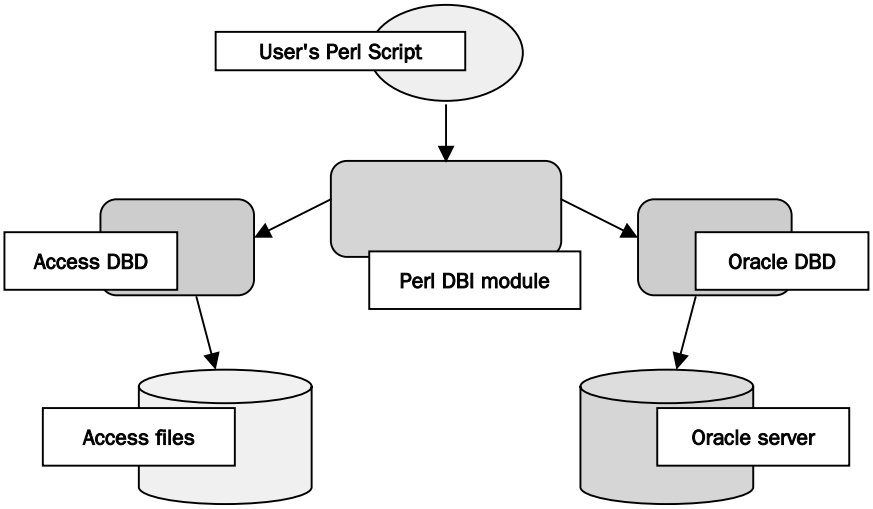


Figure 1: DBI provides an abstract interface for specific database systems.

relatively heavyweight computations each having its own resources and address spaces. Multiple processes may execute on the same machine or on different machines across the Internet. Data exchange among processes is usually done via files, pipes (channels), or lower level sockets.

Simple Unix-style process communications can be established in Perl using file I/O operations. Processes are given filehandles, and communication paths are established with the `open` statement using a pipe symbol "|" on the command the process will execute. To read from a running program, for example, the pipe goes at the end:

```
$pid = open(DATAGEN, "ls -lrt |") || die
  "Couldn't fork: $!\n";
while (<DATAGEN>) {
    print;
}
close(DATAGEN) || die "Couldn't close: $!\n";
```

This program creates a process that executes the Unix "`ls`" command with arguments "`-lrt`" to generate a listing of the current directory. The pipe symbol tells Perl that the *open* is specifying a process command to run instead of a simple file name. To write to a process, the pipe goes at the beginning of the command:

```
$pid = open(DATASINK, "|  myProg args") ||
  die "Couldn't fork: $!\n";
print DATASINK "some data for you\n";
close(DATASINK) || die "Couldn't close:
  $!\n";
```

A script can use *pipe* and *fork* to create two related processes that communicate, with better control than can be had from *open*, *system*, and *backticks*:

```
pipe(INFROM, OUTTO); # opens connected
  filehandles
if (fork) {# both processes share all open
  filehandles
    # run parent code for writing
    close(INFROM);
    # now the writer code...
} else {
    # run child code for reading
    close(OUTTO);
    # now the reader code...
}
```

For more complicated situations, such as reading and writing to the same executing process, the previous methods are not sufficient. There is a special forking form of *open* that can be used. However, using the *IPC::Open2* module is a better approach:

```
use IPC::Open2;
open2(*READFROM, *WRITETO, "myProg arg1
  arg2");
print WRITETO "here's your input\n";
$output = <READFROM>;
    #  etc...
close(WRITETO);
close(READFROM);
```

Here the program "myProg" is executed as a process, using the supplied arguments, and the filehandles READFROM and WRITETO are connected to its standard input and output respectively so the Perl script can exchange data with it.

Module "Socket"

Even finer grained control over processes can be obtained if the programmer is willing to program lower into the operating system. Sockets are the underlying technical mechanism for network programming on the Internet. Perl gives access to this level with built-in functions that operate on sockets. These include

socket to assign a filehandle
bind to associate a socket with a port and address
listen to wait for activity on the server-side connection

accept to allow incoming connection on the server side
connect to establish communications with a server
recv to get data off a socket
send to put data onto a socket
close to end it all

Sockets are given filehandles on creation, so Perl functions that operate on filehandles can be used on sockets as well. Socket functions tend to need hard-coded values related to machine specifics and IP addresses, which limits portability of scripts. The Perl module *Socket* should be used in conjunction with the Perl socket functions to preload machine-specific constants and data into Perl variables.

ON BEYOND PERL

Perl was created when object-oriented (OO) programming concepts were young and less well understood than today. Early versions of Perl, in fact, did not even contain objects or references. As understanding of OO concepts has advanced, Perl has evolved to contain OO features. They work fairly well, but because they were not part of the original design rationale of Perl, many consider them less elegant and cohesive than the original set of language features.

Two more recently developed programming languages—*Python* and *Ruby*—claim Perl in their heritage, but have been designed specifically as OO programming tools. The designers of each wanted to retain the extreme convenience and flexibility of Perl's text handling and string matching, but to incorporate as well other capabilities that go beyond Perl. The results are two notations that are still largely thought of as "scripting" languages, but with more highly integrated object semantics. Following are brief discussions of each with respect to Perl.

Python

Guido van Rossum began work on the design of Python in late 1989. One of his goals for the new language was to cater to infrequent users and not just experienced ones. Infrequent users of a notation can find rich syntax (such as that of Perl) to be more burdensome than helpful. This means that Python is a *compact* language. A programmer can easily keep the entire feature set in mind without frequent references to a manual. C is famously compact in much the same way, but Perl most certainly is not. The Perl design principle of "more than one way to do it" shows up in Perl's wealth of features, notational shortcuts, and style idioms. Van Rossum also wanted a language designed from the beginning to be object-oriented and to have clear module semantics. In Python everything is an object or class including the base data types.

Python has unusual syntax for a modern programming language. Rather than being a free-form notation, in Python white space is important for defining the block structure of a script. Indentation levels serve the same purpose in Python that pairs of "{}" do in Perl, C, and others. Here, the body of the loop and the bodies of the conditional clauses are defined by vertical alignment :

```
while x < y:
    buf = fp.read(blocksize)
    if not buf: break
    conn.send(buf)
x = 3
if x == 4:
   result = x + 2
   print x
else:
   print 'Try again.'
```

Though this initially appears to be a cumbersome throwback, in many ways this makes Python easy to use. Python has a very clean and structured layout and it is very easy to follow what's going on. Perl can frequently look noisy, and new programmers particularly can have difficulty trying to understand the behavior they see from their Perl scripts. Python programmers report that after a short training period, they can produce working code about as fast as they can type, and that they begin to think of Python as *executable pseudocode*.

Python gives programmers good support for modern programming practices such as design of data structures and object-oriented programming. The compactness causes programmers to write readable, maintainable scripts by eliminating much of the cryptic notations in Perl. In Perl's original application domains, Python comes close to but rarely beats Perl for programming flexibility and speed. On the other hand, Python is proving quite usable well beyond Perl's best application domains.

Ruby

Ruby is another language that is advertised as being a natural successor to Perl. It was developed in Japan in 1993 by Yukihiro Matsumoto and began attracting a user community in the United States by the year 2000. In Japan, Ruby has overtaken Python in popularity. Like Python, Ruby is open sourced and so is easy for others to extend, correct, or modify.

Matsumoto was looking to design an object oriented scripting language that did not have the messy "Swiss Army chainsaw" aspects of Perl, but he considered Python to be not object oriented enough. Ruby has retained many of the text manipulation and string matching features of Perl that Python leaves out, but they are elevated to the class level (e.g., regular expressions are classes). Even operators on the base types are methods of the base classes. In addition to classes, Ruby allows metaclass reasoning, allowing scripts to understand and exploit the dynamic structure of objects at runtime. Ruby is best thought of as having modern object semantics, as Python does, but also retaining more of the Perl features and syntax than Python does.

GLOSSARY

CGI Common gateway interface, standard for Web server scripting
CPAN Comprehensive Perl Archive Network, repository for useful Perl code and documentation

Hash Associate array, a collection of data items indexed by string (scalar) values
HTTP Hypertext transfer protocol, encoding standard for Web transactions
HTML Hypertext markup language, language for specifying Web page content
IPC Interprocess communication
Linux Popular open source version of Unix for PCs
Pipe Unix name for a communication channel between processes
Regular expression formal language for specifiying complex patterns and strings of characters
Scalar Singular data value such as integer, string, boolean, real
Unix Popular operating system developed at Bell Labs and U.C. Berkeley in the late 1970s

CROSS REFERENCES

See *Common Gateway Interface (CGI) Scripts; HTML/XHTML (HyperText Markup Language/Extensible HyperText Markup Language); Linux Operating System; Unix ' Operating System.*

FURTHER READING

To learn programming using Perl, consult this text:

Johnson, A. L. (1999). *Elements of Programming with Perl.* Indianapolis, IN: Manning Publications.

These books give more details on the Perl language for readers who understand programming:

Chapman, N. (1997). *Perl: The programmers companion*. New York: Wiley.
Christiansen, T., & Torkington, N. (1998). *Perl cookbook: Solutions and examples for Perl programmers*. Sebastopol, CA: O'Reilly and Associates.
Schwartz, R., & Phoenix, T. (2001). *Learning Perl: Making easy things easy and hard things possible* (3rd Ed.). Sebastopol, CA: O'Reilly and Associates.
Siever, E., Spainhour, S., & Patwardhan, N. (2002). *Perl in a nutshell*. Sebastopol, CA: O'Reilly and Associates.
Vromans, J. (2002). *Perl pocket reference* (4th Ed.). Sebastopol, CA: O'Reilly and Associates.
Wall, L., Cristiansen, T., & Orwant, J. (2000). *Programming Perl* (3rd Ed.). Sebastopol, CA: O'Reilly and Associates.

These texts give detailed, "under the covers" explanations of the advanced features in Perl:

Conway, D. (1999). *Object oriented Perl*. Greenwich, CT: Manning Publications.
Friedl, J. E. F. (2002). *Mastering regular expressions* (2nd Ed.). Sebastopol, CA: O'Reilly and Associates.
Srinivasan, S. (1997). *Advanced Perl programming.* Sebastopol, CA: O'Reilly and Associates.

These references show how to apply Perl and Perl modules in specific application domains:

Stein, L., & MacEachern, D. (1999). *Writing Apache modules with Perl and C*. Sebastopol, CA: O'Reilly and Associates.
Burke, S. M. (2002). *Perl and LWP.* Sebastopol, CA: O'Reilly and Associates.
Guelich, S., Gundavaram, S., & Birznieks, G. (2000). *CGI programming with Perl* (2nd Ed.). Sebastopol, CA: O'Reilly and Associates.
Ray, E. T., & McIntosh, J. (2002). *Perl and XML*. Sebastopol, CA: O'Reilly and Associates..
Wright, M. (1997). *CGI/Perl cookbook*. New York: Wiley Computer Publishing.

These references give more information on languages with some similarity to Perl:

Beazley, D. (2001). *Python essential reference* (2nd Ed.). Indianapolis, IN: New Riders Publishing.
Harms, D., & McDonald, K. (1999). *The quick Python book*. Greenwich, CT: Manning Publications.
Lutz, M., & Ascher, D. (1999). *Learning Python*. Sebastopol, CA: O'Reilly and Associates.
Thomas, D., & Hunt, A. (2000). *Programming Ruby: The pragmatic programmer's guide*. Reading, MA: Addison Wesley Longman. Retrieved 2002 from http://www.rubycentral.com/book/

These Web sites contain extensive information about the Perl language definition and standard, tutorials on programming in Perl, example programs and useful scripts, libraries, and upcoming conferences:

http://www.perl.com/main perl commercial distribution site
http://cpan.org/perl archives
http://use.perl.org/news clearinghouse
http://history.perl.org/PerlTimeline.html specific development timeline
http://www.perl.org/Perl mongers
http://dev.perl.org/perl6/ latest in development of Perl 6
http://www.activestate.com/ Perl implementations for Windows platforms
http://history.perl.org/perl through the ages

Finally, there are the Perl newsgroups. Use any news reader to access these groups. These provide discussions about Perl and a place to ask and answer questions.

comp.lang.perl.misc: The Perl language in general
comp.lang.perl.announce: Announcements about Perl (moderated)

Personalization and Customization Technologies

Sviatoslav Braynov, *State University of New York at Buffalo*

INTRODUCTION

Personalization and customization are considered increasingly important elements of Web applications. The terms usually refer to using information about a user (be it a customer, a Web site visitor, an individual, or a group) to better design products and services tailored to that user. One way to define personalization is by describing how it is implemented and used.

> Personalization is a toolbox of technologies and application features used in the design of an end-user experience. Features classified as personalization are wide-ranging, from simple display of the end-user's name on a Web page, to complex catalog navigation and product customization based on deep models of users' needs and behaviors. (Kramer, Noronha, & Vergo, 2000)

The Personalization Consortium (Personalization Consortium, n.d.), an international advocacy group formed to promote the development and use of personalization technology on the Web, offers the following definition:

> Personalization is the combined use of technology and customer information to tailor electronic commerce interactions between a business and each individual customer.

Personalization usually means gathering and storing information about Web site visitors and analyzing this information in order to deliver the right content in a user-preferred form and layout. Personalization helps to increase customer satisfaction, promote customer loyalty by establishing a one-to-one relationship between a Web site and its visitor, and increase sales by providing products and services tailored to customers' individual needs. The goal of personalization is to better serve the customer by anticipating his needs; it customizes services and products, and establishes a long-term relationship encouraging the customer to return for subsequent visits.

Although both customization and personalization refer to the delivery of information, products, or services tailored to users' needs, the two notions differ in several respects. Customization is usually used to describe the interface attributes that are user-controlled. That is, the user is in control and is able to configure a Web site, a product, or a service according to his or her preferences and requirements. The system is almost passive and provides only a means by which the actual configuration is done. Customization is usually done manually by the user, according to his preferences. An example of customization is My Yahoo (http://my.yahoo.com), shown in Figure 1. In My Yahoo a user can customize content by selecting from various modules (portfolios, company news, weather, currency converter, market summary, etc.) and placing them on a customized Web page, which is updated periodically. In this case, the locus of control lies with the user, who manually selects the modules on the page. In contrast, personalization is automatically performed by a Web site based on the history of previous interactions with the user, on the user's profile, or on the profiles of like-minded users. For example, Amazon.com recommends items to a user and creates personalized Web pages based on the user's navigation and purchase history.

USER PROFILING

Factual and Behavioral Profiles

Personalization requires some information about the user's preferences, needs, goals, and expectations. Information that describes a particular user is called a user profile. Adomavicius and Tuzhilin (1999) consider

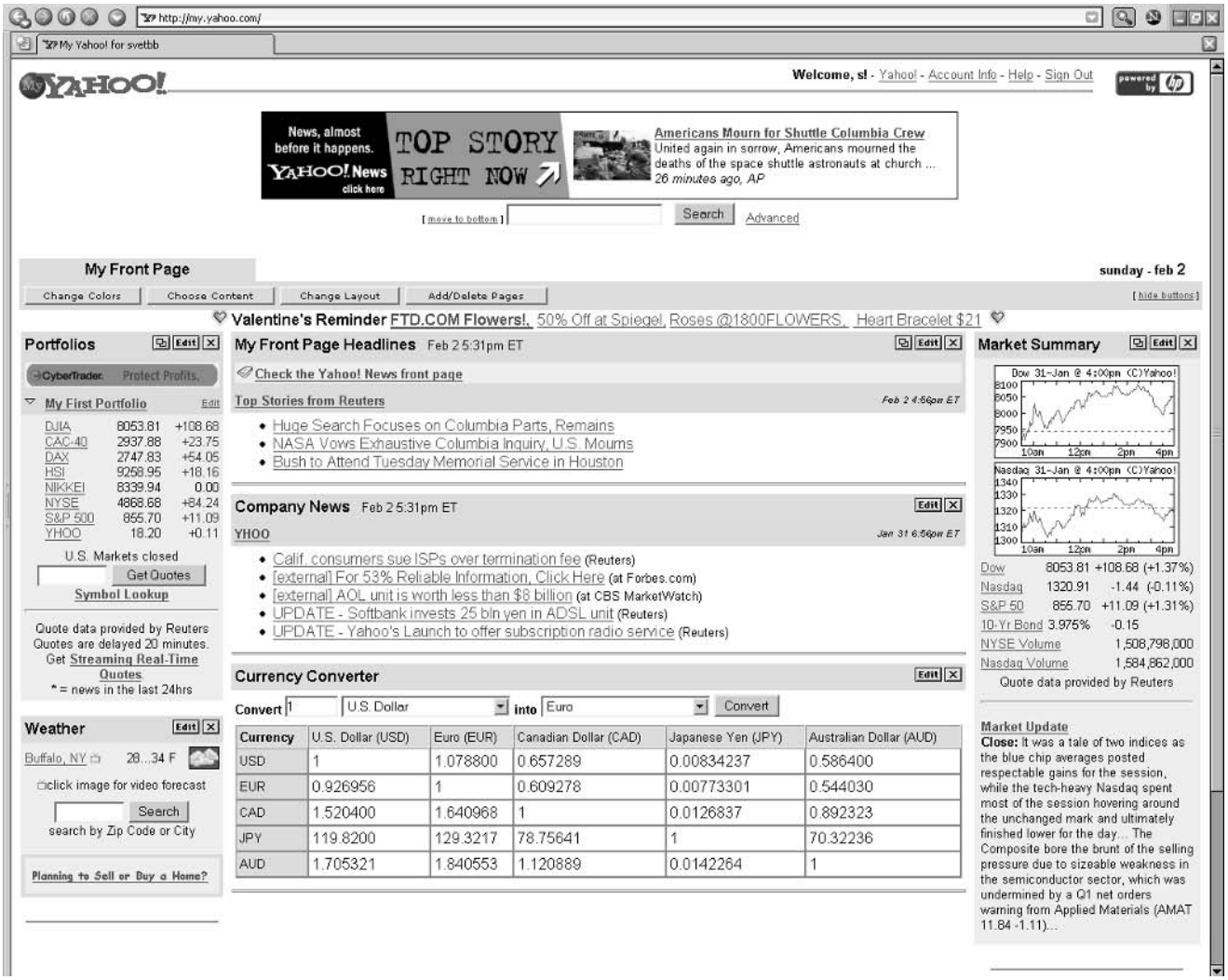


Figure 1: A customized screen of MyYahoo.

two major components of a user profile: behavioral and factual. The factual component contains demographic and transactional information such as age, income, educational level, and favorite brands. Engage Technologies (http:// www.engage.com), for example, sells software that helps companies gather and use factual profiles. The behavioral component contains information about the online activities of the customer. It is usually stored in different forms such as logic-based descriptions, classification rules, and attribute–value pairs. The most common representation of behavioral information is association rules. Here is an example of an association rule: "When shopping on weekends, consumer X usually spends more than $100 on groceries" (Adomavicius & Tuzhilin, 1999). The rules can be defined by a human expert or extracted from transactional data using data mining methods. Broad Vision (http://www.broadvision.com) and Art Technology Group (http://www.atg.com), among others, sell software that helps users build and use rule-based profiles.

The rule-based profile-building process usually consists of two main steps: rule discovery and rule validation. Various data mining algorithms such as Apriori (Agrawal & Srikant, 1994) and FP-Growth (Han, Pei, Yin, & Mao, in press) can be used for rule discovery. A special type of association rules, profile association rules, has been proposed by Agrawal, Sun, and Yu (1998). A profile association rule is one in which the left-hand side consists of customer profile information (age, salary, education, etc.), and the right-hand side of customer behavior information (buying beer, using coupons, etc.). Agrawal et al. (1998) proposed a multidimensional indexing structure and an algorithm for mining profile association rules.

One of the problems with many rule discovery methods is the large number of rules generated, many of which, although statistically acceptable, are spurious, irrelevant, or trivial. Post-analysis is usually used to filter out irrelevant and spurious rules. Several data mining systems perform rule validation by letting a domain expert inspect the rules on a one-by-one basis and reject unacceptable rules. Such an approach is not scalable to large numbers of rules and customer profiles. To solve the problem, Adomavicius and Tuzhilin (2001) proposed collective rule validation. Rules are collected in a single set to which several rule validation operators are applied iteratively. Because many users share identical or similar rules, those can be validated

together. Collective rule validation allows a human expert to reject or accept a large number of rules at once, thereby reducing validation effort.

Explicit Versus Implicit Profiling

Data for user profiling can be collected implicitly or explicitly. Explicit collection usually requires the user's active participation, thereby allowing the user to control the information in his profile. Explicit profiling can take different forms. The user may fill out a form, take part in a survey, fill out a questionnaire, submit personal information at the time of registration, provide a ranking or rating of products, etc. This method has the advantage of letting the customers tell a Web site directly what they need and how they need it.

Implicit profiling does not require the user's input and is usually performed behind the scenes. amazon.com, for example, keeps track of each customer's purchasing history and recommends specific purchases. Implicit profiling usually means tracking and monitoring users' behavior in order to identify browsing or buying patterns in customers' behavior. In many cases, tracking is performed without users' consent and remains transparent to users. Implicit data could be collected on the client or on the server side. Server-side data include automatically generated data in server access logs, referrer logs, agent logs, etc. Client-side data could include cookies, mouse or keyboard tracking, etc. Other sources of customer data are transaction databases, pre-sale and after-sale support data, and demographic information. Such data could be dynamically collected by a Web site or purchased from third parties. In many cases data are stored in different formats in multiple, disparate databases.

Implicit profiling removes the burden associated with providing personal information from the user. Instead of relying on the user's input, the system tries to collect relevant data and infer user-specific information. Although less intrusive, implicit profiling may raise several privacy concerns.

User profiles and their components can be further classified into static and dynamic, and individual and aggregated (group profiles). A profile is static when it changes seldom or never (for example, demographic information). If customer preferences tend to change over time, dynamic profiles can be used. Such profiles are periodically updated to reflect changes in consumer behavior.

OVERVIEW OF FILTERING TECHNOLOGIES

Although necessary, user profile management (creating, updating, and maintaining user profiles) is not sufficient for providing personalized services. Information in user profiles has to be analyzed in order to infer users' needs and preferences. In this section we will briefly explain the most popular personalization techniques: rule-based filtering, collaborative filtering, and content-based filtering. All these techniques are used to predict customers' interests and make recommendations.

Rule-Based Filtering

Association rule mining looks for items that tend to appear together in a given data set. Items could refer to different things in different contexts. They can be products bought by a customer, Web pages visited by a user, etc. To introduce association rules formally, we need the following notation. Let $\mathbf{I}$ denote the set of all items. A transaction $\mathbf{T}$ is defined as a set of items bought together ($\mathbf{T} \subseteq \mathbf{I}$). The set of all transactions is denoted by $\mathbf{D}$. Then, an association rule is defined as an implication between itemsets $\mathbf{A}$ and $\mathbf{B}$, denoted by

$$\mathbf{A} \Rightarrow \mathbf{B},$$

where $\mathbf{A} \subseteq \mathbf{I}$, $\mathbf{B} \subseteq \mathbf{I}$, and $\mathbf{A} \cap \mathbf{B} = \emptyset$. An association rule indicates that the presence of items in $\mathbf{A}$ in a transaction implies the presence of items in $\mathbf{B}$. For example, according to the following association rule, visitors who look at Web pages $\mathbf{X}$ and $\mathbf{Y}$ also look at Web page $\mathbf{Z}$:

look(Visitor, **X**) and look(Visitor, **Y**)
⇒ look (Visitor, **Z**).

Rules can associate items or customers. For example, the following rule associates items,

buys(Customer_1, **X**) and buys(Customer_1, **Y**)
⇒ buys(Customer_1, **Z**),

and the next rule associates customers,

buys(Customer_1, **X**) and buys(Customer_2, **X**)
⇒ buys(Customer_3, **X**).

Rule-based filtering is based on the following idea. If the behavioral pattern of a customer matches the left-hand side of a rule, then the right-hand side can be used for recommendation or prediction.

Two measures are used to indicate the strength of an association rule: *support* and *confidence*. The support of the rule $\mathbf{A} \Rightarrow \mathbf{B}$ is the fraction of the transactions containing both $\mathbf{A}$ and $\mathbf{B}$, i.e., $|\mathbf{A} \cup \mathbf{B}|/|\mathbf{D}|$. The confidence of the rule $\mathbf{A} \Rightarrow \mathbf{B}$ is the fraction of the transactions containing $\mathbf{A}$ which also contain $\mathbf{B}$, i.e., $|\mathbf{A} \cup \mathbf{B}|/|\mathbf{A}|$. Because a large number of association rules can be generated from large transaction databases, weak and nonsignificant associations have to be filtered out. To eliminate spurious associations, minimum support and minimum confidence can be used. That is, all rules that do not meet the minimum support and minimum confidence are eliminated.

An efficient algorithm for association rule mining is frequent pattern growth (FP-growth). The algorithm uses a divide-and-conquer strategy and compresses the database representing frequent items into a frequent-pattern tree (Han et al., in press).

Collaborative Filtering

Collaborative filtering (CF) was one of the earliest recommendation technologies. CF is used to make a recommendation to a user by finding a set of users, called a neighborhood, that have tastes similar to those of the target user.

Products that the neighbors like are then recommended to the target user. In other words, CF is based on the idea that people who agreed on their decisions in the past are likely to agree in the future. The process of CF consists of the following three steps: representing products and their rankings, forming a neighborhood, and generating recommendations.

During the representation stage, a customer–product matrix is created, consisting of ratings given by all customers to all products. The customer–product matrix is usually extremely large and sparse. It is large because most online stores offer large product sets ranging into millions of products. The sparseness results from the fact that each customer has usually purchased or evaluated only a small subset of the products. To reduce the dimensionality of the customer–product matrix, different dimensionality reducing methods can be applied, such as latent semantic indexing and term clustering.

The neighborhood formation stage is based on computing the similarities between customers in order to group like-minded customers into one neighborhood. The similarity between customers is usually measured by either a correlation or a cosine measure. After the proximity between customers is computed, a neighborhood is formed using clustering algorithms.

The final step of CF is to generate the top-*N* recommendations from the neighborhood of customers. Recommendations could be generated using the most-frequent-item technique, which looks into a neighborhood of customers and sorts all products according to their frequency. The *N* most frequently purchased products are returned as a recommendation. In other words, these are the *N* products most frequently purchased by customers with similar tastes or interests. Another recommendation technique is based on association rules. It finds all rules supported by a customer, i.e., the rules that have the customer on their left hand side, and returns the products from the right hand side of the rule.

Content-Based Filtering

Another recommendation technique is content-based recommendation (Balabanovic & Shoham, 1997). Although collaborative filtering identifies customers whose tastes are similar to those of the target customer, content-based recommendation identifies items similar to those the target customer has liked or has purchased in the past. Content-based recommendation has its roots in information retrieval (Baeza-Yates & Ribeiro-Neto, 1999). For example, a text document is recommended based on a comparison between the content of the document and a user profile. The comparison is usually performed using vectors of words and their relative weights. In some cases, the user is asked for feedback after the document has been shown to him. If the user likes the recommendation, the weights of the words extracted from the document are increased. This process is called relevance feedback.

However, content-based recommender systems have several shortcomings. First, content-based recommendation systems cannot perform in domains where there is no content associated with items, or where the content is difficult to analyze. For example, it is very difficult to apply content-based recommendation systems to product catalogs based solely on pictorial information. Second, only a very shallow analysis of very restricted content types is usually performed. To overcome these problems a new hybrid recommendation technique called content-boosted collaborative filtering has been proposed (Melville, Mooney, & Nagarajan, 2002). The technique uses a content-based predictor to enhance existing user data and then provides a recommendation using collaborative filtering.

In general, both content-based and collaborative filtering rely significantly on user input, which may be subjective, inaccurate, and prone to bias. In many domains, users' ratings may not be available or may be difficult to obtain. In addition, user profiles are usually static and may become quickly outdated.

WEB USAGE ANALYSIS FOR PERSONALIZATION

Some problems of collaborative and content-based filtering can be solved by Web usage analysis. Web usage analysis studies how Web sites are used by visitors in general and by each user in particular. Web usage analysis includes statistics such as page access frequency, common traversal paths through a Web site, session length, and top exit pages. Usage information can be stored in user profiles for improving the interaction with visitors. Web usage analysis is usually performed using various data mining techniques such as association rule generation and clustering.

Web Usage Data

Web usage data can be collected at the server side, the client side, or proxy servers or obtained from corporate databases. Most of the data comes from the server log files. Every time a user requests a Web site, the Web server enters a record of the transaction in a log file. Records are written in a format known as the common log file format (CLF), which has been standardized by the World Wide Web Consortium (W3C). The most useful fields of a CLF record are the IP address of the host computer requesting a page, the HTTP request method, the time of the transaction, and the referrer site visited before the current page.

Although server log files are rich in information, data are stored at a very detailed level, which makes them difficult for human beings to understand. In addition, the size of log files may be extremely large, ranging into gigabytes per day. Another problem with server log files is the information loss caused by caching. In order to improve performance most Web browsers cache requested pages on the user's computer. As a result, when a user returns to a previously requested page, the cached page is displayed, leaving no trace in the server log file. Caching could be done at local hosts and proxy servers.

Web usage data can also be collected by means of cookies containing state-related information, such as user ID, passwords, shopping cart, purchase history, customer preferences, etc. According to the W3C cookies are "the data sent by a Web server to a Web client, stored locally by the client and sent back to the server on subsequent

requests." Cookies help keep track of several visits by the same customer in order to build his profile. Some marketing networks, such as DoubleClick, use cookies to track customers across many Web sites.

Users can and often do disable cookies by changing the configuration parameters of their Web browsers. Another way to cope with marketing networks' cookies is by regularly checking the cookie files and deleting them. Some utilities, such as CookieCop, let users automatically accept or reject certain cookies. The program runs as a proxy server and monitors all cookie-related events.

It should be pointed out, however, that rejecting cookies may disable e-commerce transactions. Many retail Web sites, for example, use cookies for shopping-cart implementation, user identification, passwords, etc. Rejecting cookies may also cause problems to companies attempting customization and personalization. In general, there is a tradeoff between privacy and personalization. The more information a user reveals, the more personalized services he obtains.

Other tracking devices, currently producing much controversy, are Web bugs, or clear GIFs. A Web bug is a hidden (or very small) image in a Web page that activates a third-party spying device without being noticed by the Web page visitors. Web bugs are usually used to track users' purchasing and browsing habits.

Mechanisms for User Identification

Server log data contains information about all users visiting a Web site. To associate the data with a particular user, user identification is performed. The simplest form of user identification is user registration, in which the user is usually asked to fill out a questionnaire. Registration has the advantage of being able to collect rich demographic information, which usually does not exist in servers' logs. However, due to privacy concerns, many users choose not to browse sites requiring registration, or may provide false or incomplete information.

Another method for user identification is based on log file analysis. Log-based user identification is performed by partitioning the server log into a set of entries belonging to the same user. Accurate server log partitioning, however, is not always feasible due to rotating IP addresses at ISPs, missing reference due to local or proxy server cashing, anonymizers, etc. For example, many users can be mistakenly classified as a single user if they use a common ISP and have the same IP address. Several heuristics can be used to differentiate users sharing an IP address (Pirolli, Pitkow, & Rao, 1996). One may look for changes in the browser or operating system in the server log file. Because a user is expected to keep the same browser or operating system during his visit to a Web site, a change could mean that another visitor (with a different browser or operating system) uses the same IP address.

Another technique for user identification uses software agents loaded into browsers, which send back data. Due to privacy concerns, however, such agents are very likely to be rejected by users.

The most reliable mechanisms for automatic user identification are based on cookies. Whenever a browser contacts a Web site, it will automatically return all cookies associated with the Web site. Cookie-based user identification is reliable if the user launches each URL request from the same browser.

Another problem with user identification is that Web sites typically deal with both direct and indirect users (Ardissono & Goy, 2001). A customer is an indirect user if he visits a Web site on behalf of someone else. For example, a user may visit a Web store in order to buy a gift for a relative. In this case, the Web store must personalize gift suggestions and recommendations to the preferences of the intended beneficiary (the relative) and not to the preferences of the visitor. To overcome this problem, CDNOW (http://www.cdnow.com) offers a "gift advisor" which differentiates between direct and indirect users.

Session Identification

A user session consists of all activities performed by a user during a single visit to a Web site. Because a user may visit a Web site more than once, a server log may contain multiple sessions for a given user. Automatic session identification can be performed by partitioning log entries belonging to a single user into sequences of entries corresponding to different visits of the same user. Berendt, Mobasher, Spiliopoulou, and Wiltshire (2001) distinguish between time-oriented and navigation-oriented sessionizing. Time-oriented sessionizing is based on timeout. If the duration of a session or the time spent on a particular Web page exceeds some predefined threshold, it is assumed that the user has started a new session.

Navigation-based sessionizing takes into account the links between Web pages, the Web site topology, and the referrer information in a server log. A Web page P_1 is a referrer to another page P_2 if the URL request for P_2 was issued by P_1, i.e., the user came to P_2 by clicking on a link on P_1. A common referrer heuristic is based on the assumption that a user starts a new session whenever he uses a referrer different from or not accessible from previously visited pages. For example, if a user comes to page P_2 with a referrer page P_1 and P_2 is not accessible from P_1 given the Web site topology, then it is reasonable to assume that the user has started a new session. This heuristic, however, fails when the user uses the "Back" button or chooses a recent link kept by the browser.

Clickstream Analysis

Clickstream analysis is a special type of Web usage mining that provides information essential to understanding users' behavior. The concept of clickstream usually refers to a visitor's path through a Web site. It contains the sequence of actions entered as mouse clicks, keystrokes, and server responses as the visitor navigates through a Web site. Clickstream data can be obtained from a Web server log file, commerce server database, or from client-side tracking application.

Most efforts in Web usage analysis are focused on discovering users' access patterns. Understanding users' navigation through a Web site can help provide customized content and structure tailored to their individual needs. Chen, Park, and Yu (1996) proposed an algorithm for mining maximal forward reference, where forward reference is defined as a sequence of pages requested by a user up to

the last page before backtracking occurs. The algorithm first converts server log data into a set of maximal forward references and then determines frequent traversal patterns, i.e., maximal forward references that occur frequently.

Another approach is taken in the Web Utilization Miner (WUM) (Spiliopoulou & Faulstich, 1999). The authors of WUM not only identify sequences of frequently accessed pages but also find less frequent paths having structural or statistical properties of interest. In WUM the path followed by a user is called a trail. Because many users can display similar navigation patterns, users' trails are aggregated into a tree by merging trails sharing a prefix. The aggregate tree can be subsequently used for predicting user's behavior.

Markov models have also been used for modeling users' browsing behavior (Deshpande & Karypis, 2001). Markov models predict the Web page accessed by a user given the sequence of Web pages previously visited. Such models have proved to display high predictive accuracy. On the other hand, they have high state-space complexity, which significantly limits the scope of their applicability.

Mobasher, Dai, Luo, and Nakagawa (2002) proposed an aggregate profiling algorithm based on clustering transactions (PACT). User's transactions are represented as multidimensional space vectors of page views. The vectors are grouped into clusters, each of them representing a set of users with similar navigational patterns. Subsequently, every cluster is associated with a single point (the cluster's centroid) representing an aggregate profile of all users in that cluster. A new user activity is matched against aggregate profiles and items are recommended based on the degree of matching.

INTELLIGENT AGENTS FOR PERSONALIZATION

Intelligent agent technology provides a useful mechanism for personalization and customization. The term intelligent agent has been used in different meaning by different authors. By agent we refer here to a software program acting on behalf of its user and having the following properties (Weiss, 1999):

Autonomy: agents operate without the direct intervention of their user.

Social ability: agents interact with other agents (including humans) via agent-communication language.

Reactivity: agents perceive and adapt to a dynamically changing environment.

Proactiveness: agents do not simply act in response to their environment; they are able to exhibit goal-directed behavior.

Rationality: every agent has a representation of its user's preferences and tries to satisfy them in the best possible way.

Intelligent agents come in different types. *Internet agents* help the user collect, manipulate, and analyze information. Some of them are embedded within an Internet browser and help the user navigate through a Web site. *Interface agents* act like personal assistants collaborating with the user. Interface agents monitor, observe, and learn from the user's actions. Some interface agents can assume the form of a synthetic character; other can model users' emotions or chat with the user, using natural language. *Collaborative agents* act as a team to achieve some common goal. *Mobile agents* can roam the Internet and interact with foreign hosts and other agents on behalf of their users.

In other words, intelligent agents can be viewed as proxies of human users, capable of learning and reasoning about users' preferences. WebWatcher (Joachims, Freitag, & Mitchell, 1997) was one of the first software agents to assist users browsing the Web. It guides users through a Web site by trying to predict the paths they will follow based on the navigation history of previous users of the Web site. WebWatcher believes that a particular hyperlink is likely to be followed by a user if like-minded visitors previously followed it. WebWatcher suggests a link based on current user, user's interest, and a Web page. The user's interest is represented by a set of keywords and a hyperlink is represented by a feature vector. When a new user enters a Web page, WebWatcher compares the current user's interest with the descriptions of the links on the page and suggests the links correlated with the user interest. WebWatcher also uses reinforcement learning to learn how to navigate through a Web site. The learning is based on positive reinforcement WebWatcher receives whenever it chooses a link that fits the user's interests.

Letizia (Lieberman, 1997) is another software agent for client-side personalization. It learns a model of its user by observing his or her behavior. The user model consists of a list of weighted keywords representing the user's interests. Letizia explores the Web ahead of its user and recommends potential links of interest. It records every choice of the user on a Web page and takes the act of viewing a Web page as a positive feedback (evidence of interest). Letizia tries to incrementally learn the user's interest in a page by observing the choices he or she makes. This is an example of unobtrusive personalization, which does not require explicit user interaction: the user is not asked to explicitly rank or evaluate Web pages.

LOCATION-BASED PERSONALIZATION

The information used for personalization may range from a history of past purchases and browsing behavior to explicitly provided user preferences. The rapid growth of wireless networks and mobile commerce provide new opportunities for personalization by offering more user-specific information such as geographic location, date, time, and travel direction. Handheld devices, for example, allow customers to receive personalized content and recommendations on the move, at home, and at work.

One of the most promising technologies is location-based services (LBS), which allows business to identify a user's location and offer context-dependent services. LBS holds the potential to significantly improve CRM, wireless marketing, and emergency services.

In October 2000, Ericsson, Motorola, and Nokia founded the location interoperability forum (LIF) established to provide location-based services. The forum

aims at providing a common device location standard to achieve global interoperability between location-based services regardless of their underlying technologies. At present, there is a wide range of location identification technologies. One method, for example, involves the cell ID number, which identifies a cellular device to a network. Other methods (Deitel, Deitel, Nieto, & Steinbuhler, 2002) are triangulation (used by several satellites in GPS), cell of origin (a cellular phone is located by a nearby tower), angle of arrival (several towers measure the angles from which a cellular phone's signals are received), and observed time difference (the travel time between a cellular phone and multiple towers is measured to determine the phone's location).

Location-based services allow content providers to offer personalized services based on customers' geographic position. Mobile users can receive local weather reports, news, travel reports, traffic information, maps, hotels, restaurant information, etc. For example, Go2 Systems (www.go2online.com) provides a mobile Yellow Pages directory based on users' location. The directory allows users to get directions to various nearby services such as entertainment, real estate, finance, recreation, government, and travel.

One problem with LBS is that the small screen and the limited capabilities of wireless devices can reduce the level of personalization. For example, it is often impractical to use Web-based registration forms or questionnaires for explicit personalization. Billsus, Brunk, Evans, Gladish, and Pazzani (2002) report that only 2–5% of wireless users customize their interfaces, due to technical problems or poor content management. Another problem is complex navigation and the structure of WAP (wireless application protocol) sites. Each WAP site consists of multiple decks, each of them containing one or more cards. Hypertext links can be made between cards in the same or in different decks. As a result, users must make too many selections and move through too many cards in order to achieve their goals. In addition, the limited processing power and slow network access for many mobile devices lead to extended response times.

To improve Web site navigation for wireless devices, Anderson, Domingos, and Weld (2001) proposed a personalization algorithm, MinPath, which automatically suggests useful shortcut links in real time. MinPath finds shortcuts by using a learned model of Web visitor behavior to estimate the savings of shortcut links and to suggest only the few best links. The algorithm takes into account the value a visitor receives from viewing a page and the effort required to reach that page.

PREFERENCE MODELING

User preferences play an important role in user modeling, personalization, and customization. According to the decision-theoretic tradition, human preferences are modeled as a binary relation **R** over a set of possible alternatives such as products and services, information content, layout, and interaction style. A preference relation **R** holds between two alternatives (or choice options) **X** and **Y** (i.e., **R**(**X**, **Y**)) if **X** is more preferred to **Y**. That is, the user will chose **X** when he or she faces a choice between **X** and **Y**. The indifference between **X** and **Y** could be represented as **not**(**R**(**X**,**Y**)) **and not**(**R**(**Y**,**X**)). It has been proved (Fishburn, 1970) that, under certain conditions, a preference relation can be represented by an order-preserving numeric function **U** called a utility function. In other words, alternative **X** is preferred to alternative **Y** if and only if the utility of **X** is greater than the utility of **Y**, (**U**(**X**) > **U**(**Y**)). Knowing a user's utility function allows a system to offer its customers those products, services, or information that maximize the customers' utility. The problem of preference modeling has been approached relatively recently in the computer science community. Usually users' preferences are represented based on ad hoc approaches such as attribute–value pairs and association rules. It is still an open question how to represent user preferences in a computationally tractable way and how to reason with incomplete or inaccurate preferences. In this section we discuss three important preference problems: how to elicit user's preferences, how to reason with conditional preferences, and how to take advantage of users' preferences in personalizing access to databases.

Preference Elicitation

The process of preference elicitation consists of finding a user's preferences for a well-defined set of choices (for example, products). In general, preference elicitation could be performed by interviewing or observing user's behavior. Various methods for preference elicitation have been proposed. Most of them assume that consumer preferences are additive functions over different attributes or decision objectives. That is, a user multiattribute utility function is a weighted sum of single-attribute utility functions,

$$U(x_1, \ldots, x_n) = \sum w_i \, U_i(x_i),$$

where $U(x_1, \ldots, x_n)$ is the user utility function, $U_i(x_i)$ are single-attribute utility functions, and w_i are their weights. The intuition behind this assumption is that it may be natural to think in terms of utility for each attribute and then combine these utilities into an overall multiattribute utility function.

The analytic hierarchy process (AHP) (Saaty, 1980) is a common method for discovering attribute weights w_i. AHP is carried out in two steps. In the first step, an attribute hierarchy is set up. In the second step, the user is asked to compare attributes sharing a parent in the hierarchy. Pairwise attribute comparisons determine the relative importance of each attribute with respect to the attribute on the level immediately above it. The strength of the comparison is measured on a ratio scale. The comparisons are used to build a reciprocal matrix, which is subsequently used to derive the relative weights w_i for the overall utility function.

Another method for preference elicitation is multiattribute conjoint analysis (Luce, 1977). Attribute values are usually discretized and every combination of discrete attribute levels is ranked by the user. The rank is then used as its utility value. The coefficients w_i in the overall utility function are derived using statistical methods such as regression analysis.

Direct approaches to preference elicitation, however, are generally not feasible, due to the exponential number of comparisons and due to the complexity of the questions to be asked. In applications such as recommender systems, product configuration, and adaptive Web stores, users cannot be expected to have the patience to provide detailed preference relations or utility functions. In addition, users cannot always be expected to provide accurate and complete information about their preferences. In general, direct preference elicitation requires a significant level of introspection and imposes a significant cognitive overload on users. Instead of using direct preference elicitation, Chajewska, Getoor, Norman, and Shahar (1998) have used classification to identify a user's utility function. The authors partition users' utility functions into clusters with very similar utility functions in each cluster. Every cluster is characterized by a single reference utility function called prototype. Then for every new user the system finds the cluster to which he is most likely to belong, and uses the prototype utility function for that cluster to predict his preferences.

Reasoning with Conditional Preferences

Most of the research on preference modeling focuses on additive utility functions in which every attribute of the function is independent of other attributes. Such utility functions allow a very elegant representation in the form of a weighted sum of attribute utilities. For example, if a user's preferences for product quality **Q** are independent of his preferences for product brand **B**, then the user's utility could be represented as a weighted sum,

$$\mathbf{U}(\mathbf{Q},\mathbf{B}) = \mathbf{w}_1\mathbf{U}(\mathbf{Q}) + \mathbf{w}_2\mathbf{U}(\mathbf{B})$$

where $\mathbf{U}(\mathbf{Q},\mathbf{B})$ is the overall product utility, $\mathbf{U}(\mathbf{Q})$ is the utility for product quality, and $\mathbf{U}(\mathbf{B})$ is the utility for brand. Additive utility functions are easy to elicit and evaluate. A problem arises, however, if preferences over product attributes are not independent. For example, a preference for brand usually depends on the quality level and vice versa. To overcome this problem, Boutilier, Brafman, Hoos, and Poole (1999) proposed a model of conditional preferences. Preferences are represented as a network called a conditional preference network (CP-network) that specifies the dependence between attributes. In CP-networks a user can specify a conditional preference in the form:

If product quality is medium, then I prefer Brand_1 to Brand_2.

Boutilier et al. (1999) have also formulated and studied the product configuration problem, i.e, what is the best product configuration that satisfies a set of customer preferences. The authors have proposed several algorithms for finding an optimal product configuration for a given set of constraints. In Domshlak, Brafman, & Shimony (2001) CP-networks are used for Web-page personalization. The optimal presentation of a Web page is determined by taking into account the preferences of the Web designer, the layout constraints, and the viewer interaction with the browser. For example, the preferences of the Web page designer are represented by a CP-network and constrained optimization techniques are used to determine the optimal Web page configuration.

Preference-Based Queries

Preference modeling plays an increasingly important role in e-commerce applications where the users face the problem of information overload, i.e., how to choose from among thousands and even millions of products and product descriptions. Preference queries (Chomicki, 2002) have been proposed as a tool to help a user formulate the most appropriate queries and receive the results ranked according to his preferences. Preference queries are based on a preference operator (called winnow) which picks from a given relation the set of the most preferred tuples, according to a given preference formula.

For example, consider the instance of the book relation shown in Figure 2 (Chomicki, 2002). A user's preferences could be formulated as follows: the user prefers a book to another if and only if their ISBNs are the same and the price of the first book is lower. In this case, because the second book is preferred to the first and to the third, and there are no preferences defined between the last two books, the winnow operator will return the second, the fourth, and the fifth book. Chomicki (2002) has studied the properties of the winnow operator and has proposed several algorithms for evaluating preference queries.

APPLICATIONS

This section discusses how personalization techniques can be used to tailor the content, products, and interactions to customers' needs. Some typical personalization applications are explained, such as adaptive Web sites, recommender systems, adaptive Web stores, and customer relationship management.

Adaptive Web Sites

Web sites have been traditionally designed to fit the needs of a generic user. Adaptive Web sites, which customize content and interface to suit individual users or groups of users, provide a more effective way to interact with these users. An adaptive Web site can semiautomatically improve its organization and presentation by learning from

ISBN	Vendor	Price
0679726691	BooksForLess	14.75
0679726691	LowestPrices	13.50
0679726691	QualityBooks	18.80
0062059041	BooksForLess	7.30
0374164770	LowestPrices	21.88

Figure 2: An instance of a book relation.

visitors' access patterns (Perkowitz & Etzioni, 2000). An adaptive Web site may automatically create new pages, remove or add new links, highlight or rearrange links, reformat content, etc. In general, Web site adaptiveness could be classified as being either individual- or group-oriented. An individually adaptive Web site consists of a large number of versions—one for each individual user. Group-oriented adaptiveness targets groups of users and requires a smaller number of Web site versions—one for each group. For example, a Web site may have one view for corporate users and another view for individual users, or one view for domestic visitors and another view for international visitors.

Depending on the way it is performed, adaptation can be classified as content-based or as access-based. Content-based adaptation involves presenting and organizing Web pages according to their content. Access-based adaptation is based on the way visitors interact with a Web site. It involves tracking users' activity and personalizing content and layout according to users' navigation patterns.

The problem of designing personalized Web sites is complicated by several factors. First, different users may have different goals and needs. Second, the same user may have different goals at different times. Third, a Web site may be used in a way different from the designer's expectations. It is still not clear what kind of adaptation can be automated and what is the appropriate tradeoff between user-controlled and automatically guided navigation.

The idea of adaptive Web sites was popularized by Perkowitz and Etzioni (2000). They proposed the PageGather algorithm, which automatically generates index pages that facilitate navigation on a Web site. An index page is a page consisting of links to a set of existing but currently unlinked pages that cover a particular topic of interest. In order to find a collection of related pages on a Web site, the PageGather algorithm employs cluster mining. The algorithm takes the Web server access log, computes the co-occurrence frequencies between pages, and creates a similarity matrix. The matrix is then transformed into a similarity graph, and maximal cliques are found. Each maximal clique represents a set of pages that tend to be visited together.

Recommender Systems

Recommender systems (Schafer, Konstan, & Riedel, 2001) have been used in B-to-C e-commerce sites to make product recommendations and to provide customers with information that helps them decide which product to buy. Recommender systems provide a solution to the problem of how to choose a product in the presence of an overwhelming amount of information. Many e-commerce sites offer millions of products, and therefore, choosing a particular product requires processing large amounts of information, making a consumer choice difficult and tedious.

Recommender systems contribute to the success of e-commerce sites in three major ways (Schafer et al., 2001). First, they help to improve cross-sell. Cross-sell is usually improved by recommending additional products for the customer to buy. For example, by looking at the products in the customer's shopping cart during the checkout process, a system could recommend additional complimentary products. Second, recommender systems could help convert occasional visitors into buyers. By providing a recommendation, a retailer could deliver customized information, increase the amount of time spent on a Web site, and finally, increase the customer's willingness to buy. Third, recommender systems help build loyalty and improve customer retention. Personalized recommendations create a relationship between a customer and a Web site. The site has to invest additional resources into learning customers' preferences and needs; on their part, customers have to spend time teaching a Web site what their preferences are. Switching to a competitor's Web site becomes time-consuming and inefficient for customers who have to start again the whole process of building personalized profiles. In addition, customers tend to return to Web sites that best match their needs.

Recommender systems use different methods for suggesting products (Schafer et al., 2001). One of the most common methods is item-to-item correlation. This method relates one product to another using purchase history, attribute correlation, etc. CDNOW, for instance, suggests a group of artists with styles similar to that of the artist that the customer likes. Another recommendation method is user-to-user correlation. This method recommends products to a customer based on a correlation between that customer and other customers visiting the same Web site. A typical example is "Customers who bought" in Amazon.com. When a customer is buying or browsing a selected product, this method returns a list of products purchased by customers of the selected product. Other recommendation methods include statistical summaries. For example, the Purchase Circles feature of Amazon.com allows customers to view the "top 10" list for a given geographic region, company, educational institution, government, or other organization.

CDNOW enables customers to set up their own music stores, based on albums and artists they like. Customers can rate albums and provide feedback on albums they have purchased. The system keeps track of previously purchased albums and predicts six albums the customer might like based on what is already owned.

A recent study by Haubl and Murray (2001) shows that recommendation algorithms have the potential to manipulate and influence user preferences in a systematic fashion. The authors performed a controlled experiment showing that the inclusion of a product feature in a recommendation renders the feature more prominent in customers' purchase decisions.

Adaptive Web Stores

Adaptive Web stores are a special type of adaptive Web sites that can use a customer profile to suggest items best fitting the customer's needs. The main difference between adaptive Web stores and adaptive Web sites is that Ardissono and Goy (2001) describe a prototype toolkit (SETA) for creating adaptive Web stores. In contrast to other adaptive Web stores, SETA adapts not only item selection, but also layout and content selection to the preferences and expertise of customers. SETA chooses a catalog presentation style (item description, background colors,

Identification:
- *First name*: Joe
- *Family name*: Smith

Personal data:
- *Age*: 20
- *Gender*: male
- *Job*: student

Preferences:
- *Quality*:
  - Importance: 0.8
  - Values: low 0.1; medium 0.1; high 0.8
- *Price*:
  - Importance: 0.7
  - Values: low 0.6; medium 0.3; high 0.1
- *Ease of use*:
  - Importance: 0.7
  - Values: low 0.1; medium 0.1; high 0.8

Figure 3: An example of a user model.

font face and size) that best fits customer preferences and perceptivity.

In order to provide personalized interaction, SETA employs user's models. A user model consists of a fixed part, containing a set of domain-independent user's attributes, and a configurable part, containing the user's preferences for domain-dependent product properties. An example of a user model is given in Figure 3.

Data in a user model are represented as a list of <feature, value> pairs. The value slot represents a probability distribution over the set of possible values for a given feature. For example, the user presented in Figure 3 prefers high-quality products with probability 0.8 and medium and low quality products with probability 0.1. The importance slot describes the system's estimate of the relevance of a particular preference. Following the previous example, the user attaches an importance of 0.8 to product quality.

SETA divides all customers into groups according to the similarity of their preferences. All customers belonging to one group are described by a group profile called a stereotype. The classification into stereotypes is used to evaluate how closely a customer visiting a Web store matches a stereotypical description. The stereotype closest to a customer's properties is used for predicting his preferences, his product selection, and the interface design. Every stereotype consists of a conditional part and a prediction part. The conditional part describes the general properties of a group of customers. For example, the conditional part may assert that 30% of customers in a group are below 30 years of age, 60% of customers are between 30 and 50, and 10% are above 50 years of age. The prediction part of a stereotype is similar to a user model and describes group preferences for product properties.

When a user visits a Web store, the system tries to predict the user preferences by computing the degree of matching between the user's profile and each stereotype. The predictions of all stereotypes are merged as a weighted sum of predictions suggested by each stereotype, where the weights represent the user's degree of matching with a stereotype.

The main advantage of SETA is that the system tailors graphical design, product selection, page content and structure, and terminology to customers' receptivity, expertise, and interests. In addition, it maintains a model of each customer and of large groups of customers. On the other hand, SETA depends essentially on customer registration. The system is unable to observe customers' behavior in order to automatically build individual profiles. Another drawback of SETA is that stereotypes (group profiles) are prepared manually and, therefore, may not reflect the actual dynamic properties of customer population. An alternative approach is to use data mining techniques that automatically build and dynamically update group profiles.

Customer Relationship Management

Customer relationship management (CRM) refers to providing quality service and information by addressing customer needs, problems, and preferences. CRM can include sales tracking, transaction support, and many other activities. A CRM system usually consists of a database of customer information and tools for analyzing, aggregating, and visualizing customer information. To achieve its goal, CRM makes extensive use of personalization and customization technologies such as log-file analysis, data mining, and intelligent agents. For example, many CRM systems store and analyze consumer profiles in order to develop new products, increase product utilization, optimize delivery costs, etc.

The market for integrated solutions for online CRM is growing at a rapid pace. BroadVision (http://www.broadvision.com), for example, provides solutions for contextual personalization with its one-to-one portal and one-to-one commerce center. BroadVision combines rule-based personalization with intelligent agent matching to dynamically tailor relevant information to customers.

They also provide profile generation, session and event-based monitoring.

NetPerceptions (http://www.netperceptions.com) is another provider of CRM systems, based on collaborative filtering. NetPerceptions provide tools for one-to-one marketing and for real-time cross-sell and up-sell recommendations. The core technology is the GroupLens software for generating product recommendations.

Vignette Corporation is another provider of software for dynamic content management. Their Vignette V6 relationship manager server uses rule-based filtering, user's viewing activities, and historical data to generate customized recommendations. The system also allows business users to define their own rules for content delivery.

PERSONALIZATION AND PRIVACY

With personalized online service, concerns about privacy arise. In general, e-commerce sites must strike a difficult balance: they must recognize a returning customer without violating his privacy. According to Alan Westin (1997) privacy is "the claim of individuals, groups, or institutions to determine for themselves when, how, and to what extent information about them is communicated to others." Many personalized Web sites collect and store personal information, keep track of user online behavior, or build individual profiles without the consumer's consent. On June 13, 2000, the Federal Trade Commission (FTC) issued *Online Profiling: A Report to Congress*. The report found that many banner ads displayed on Web sites are selected and delivered by networks of advertising companies (such as 24/7 Media, AdForce, AdKnowledge, Avenue A, Burst Media, DoubleClick, Engage, and MatchLogic) without the consent and knowledge of customers. Advertising networks can track consumer behavior over large networks of interrelated Web sites and build consumer profiles. Although the profiles are usually anonymous (i.e., they are linked to a cookie or a session ID number), many advertising networks also have sociodemographic profiles (acquired from third parties) that could eventually be linked to anonymous profiles. For example, in 1999 DoubleClick purchased Abacus, a direct marketing company, with a database of over 88 million buyers profiles collected from catalog retailing. DoubleClick planned to merge that database with its own database containing clickstream browsing patterns for over 10 million Internet users.

Some privacy advocates believe that even anonymous profiles permit *digital redlining* and *weblining*. Digital redlining refers to the ability of a Web site to limit the information customers want to see to that chosen by marketers. This holds the potential of manipulating the shopping environment to the advantage of the merchant and influencing customers' purchasing decisions and buying habits. The concept of weblining refers to discriminating between customers based on their profiles and charging selected customers higher prices.

The FTC report outlined the following fair information practices:

Notice: Web sites collecting data must disclose their information practices before collecting personal information from users.

Choice: Users must be given choice with respect to whether and how personal information collected from them may be used.

Access: Users should be able to access and check the accuracy and completeness of information collected about them.

Security: Web sites collecting personal data must ensure that these data will not be used without authorization.

There are many technology-based solutions for privacy protection. One expected to have significant impact is the Platform for Privacy Preferences (P3P) proposed by the World Wide Web Consortium. P3P is designed to enable users to exercise preferences over Web sites privacy practices. It allows users to compare a Web site's privacy policy to their own standards prior to visiting a Web site or disclosing private information. P3P includes a standard vocabulary for describing privacy policies, a set of base data elements that Web sites can use in their policies, and a protocol for requesting privacy policies. Privacy policies are specified in XML format and can automatically be fetched and analyzed by a Web browser. If the Web site's policies do not agree with the user's privacy preferences, the browser can either warn the user or disable certain functionality. P3P automates privacy statement disclosure and eliminates the tedious and repetitive process of reading privacy statements.

P3P has received a lot of criticism, mostly because it does not address the issue of enforcement of privacy agreements between users and sites. P3P does not establish specific privacy standards; instead, it provides a framework on which to build privacy mechanisms.

Another type of privacy protection tools is anonymizers (Anonymizer.com, Zero Knowledge Systems, safeWeb). They serve as proxies between browsers and Web sites and hide the user's identity. Many of them offer anonymous browsing, file downloads, e-mail, etc. The main disadvantage of anonymizers is that they cannot support e-commerce transactions, which usually require the transfer of financial and personally identifiable information.

Many personalization advocates believe that the Internet will make a significant leap forward in efficiency when it will automatically recognize the digital identities of individual users. Digital identity is the codification and archiving of personally identifiable information, i.e., information from which a person can be identified (such as name, address, SSN, fingerprints, and retinal scan). Currently there are two prevailing identity services: Liberty Alliance Project Liberty 1.0 and Microsoft.Net Passport. Both of them rely on the concept of federated authentication, which allows a user to visit several Web sites with a single sign-on. For example, in Microsoft.Net Passport the user profile is stored on a Microsoft server that (with the user's approval) shares the information with participating Web sites. This raises some doubts as to whether digital identity services provide sufficient protection of privacy. Users can easily lose the ability to control how and to what extent information about them is shared with marketing firms, governmental agencies, and other third parties.

Liberty Alliance does not centralize personal information. Instead, information is distributed across several

participating companies, such as Citigroup, General Motors Corporation, and Sony Corporation, which together form the Liberty Alliance Network. Liberty Alliance also allows users to decide whether to link their accounts to different participating sites. For example, users can choose to opt out and not link their accounts to a specific site. The main difference between Microsoft Passport and Liberty Alliance is that Microsoft keeps all the data about individual users, whereas Liberty Alliance allows the data to be owned by many Web sites.

CONCLUSION

Personalization and customization are among the fastest growing segments of the Internet economy. They provide several advantages to both businesses and customers. Customers benefit from personalization by receiving customized experience, reduced information overload, and personalized products and services. Businesses benefit from the ability to learn consumers' behavior, provide one-to-one marketing, increase costumer retention, optimize product selection, and provide build-on-demand services.

Due to the large variety of personalization techniques and applications, this survey is by no means exhaustive. Recent developments in Web services, for example, offer new prospects for personalization and customization. SUN recently presented a new vision of context-sensitive Smart Web services, which can adapt their behavior to changing conditions. A smart Web service can adapt depending on a user's location or preferences.

Personalization is by no means limited to Web site data mining, machine learning, or statistical analysis. Personalization can use any technology that provides insight into customer behavior and customer preferences. It is our hope and expectation that the near future will offer new challenging technologies and that personalization will continue to be one of the most exciting fields in the modern Internet economy.

GLOSSARY

Anonymizer A proxy between a browser and a Web site which hides the user's identity.

Collaborative filtering Recommendation technology that makes a suggestion to a target customer by finding a set of customers with similar interests.

Content-based recommendation Recommendation technology that makes a suggestion to a target customer by finding items similar to those he or she has liked or has purchased in the past.

Cookie The data sent by a Web server to a Web client, stored locally by the client and sent back to the server on subsequent requests.

Customer profile A collection of data describing an individual user or a group of users.

Data mining The nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data.

Intelligent agent A software program which can act as a proxy of a human user by learning and reasoning about users' preferences.

Log file A file generated by a Web server, which keeps a record for every transaction.

Personalization Using user information to better design products and services tailored to the user.

Privacy The claim of individuals, groups, or institutions to determine for themselves when, how, and to what extent information about them is communicated to others.

User identification The process of associating Web site visits and navigation behavior with a particular user.

User session All activities performed by a user during a single visit to a Web site.

Web bug A hidden image in a Web page that activates a third-party spying device without being noticed by the Web page visitors.

CROSS REFERENCES

See *Data Mining in E-commerce; Intelligent Agents; Machine Learning and Data Mining on the Web; Rule-Based and Expert Systems*.

REFERENCES

Adomavicius, G., & Tuzhilin, A. (1999). User profiling in personalization applications through rule discovery and validation. In *Proceedings of KDD-99* (pp. 377–381). New York: ACM.

Adomavicius, G., & Tuzhilin, A. (2001). Expert-driven validation of rule-based user models in personalization applications. *Data Mining and Knowledge Discovery, 5*, 33–58.

Agrawal, R., & Srikant, R. (1994). Fast algorithms for mining association rules. In J. Bocca, M. Jarke, and C. Zaniolo (Eds.), *Proceedings of the International Conference on Very Large Data Bases (VLDB'94)* (pp. 487–499). San Francisco: Kaufmann.

Agrawal, C., Sun, Z., & Yu, P. (1998). Online generation of profile association rules. In *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining* (pp. 129–133). New York: ACM.

Anderson, C., Domingos, P., & Weld, D. (2001). Adaptive Web navigation for wireless devices. In *Proceedings of the 17th International Joint Conference on Artificial Intelligence (IJCAI)* (pp. 879–874). San Francisco: Kaufmann.

Ardissono, L., & Goy, A. (2000). Tailoring the interaction with users in Web stores. *User Modeling and User-Adapted Interaction, 10*(4), 251–303.

Balabanovic, M., & Shoham, Y. (1997). Combining content-based and collaborative recommendation. *Communications of the ACM, 40*(3), 66–72.

Baeza-Yates, R., & Ribeiro-Neto, B. (1999). *Modern information retrieval*. Reading, MA: Addison–Wesley.

Berendt, B., Mobasher, B., Spiliopoulou, M., & Wiltshire, J. (2001). Measuring the accuracy of sessionizers for Web usage analysis. In *Proceedings of the Web Mining Workshop at the SIAM International Conference on Data Mining* (pp. 7–14). University City Science Center, PA: SIAM.

Billsus, D., Brunk, C., Evans, C., Gladish, B., & Pazzani, M. (2002). Adaptive interfaces for ubiquitous Web access. *Communications of the ACM, 45*(5), 34–38.

Boutilier, C., Brafman, R., Hoos, H., & Poole, D. (1999). Reasoning with conditional ceteris paribus preference statements. In *Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence* (pp. 71–80). San Francisco: Morgan Kaufmann.

Chajewska, U., Getoor, L., Norman, J., & Shahar, Y. (1998). Utility elicitation as a classification problem. *Uncertainty in Artificial Intelligence, 14*, 79–88.

Chen, M., Park, J., & Yu, P. (1996). Data mining for path traversal patterns in a Web environment. In *Proceedings of the 16th International Conference on Distributed Computing Systems* (pp. 385–392). Austin, TX: IEEE Computer Society Press.

.Chomicki, J. (2002). Querying with intrinsic preferences. In *Proceedings of the 8th International Conference on Extending Database Technology* (pp. 34–51). New York: Springer.

Deitel, H., Deitel, P., Nieto, T., & Steinbuhler, K. (2002). *Wireless Internet and mobile business: How to program*. Upper Saddle River, NJ: Prentice Hall.

Deshpande, M., & Karypis, G. (2001). Selective Markov models for predicting Web-page accesses. In *Proceedings of the First SIAM International Conference on Data Mining*. University City Science Center, PA: SIAM.

Domshlak, C., Brafman, R., & Shimony, S. (2001). Preference-based configuration of Web page content. In *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence (IJCAI-01)* (pp. 1451–1456). San Francisco: Morgan Kaufmann.

Fishburn, P. (1970). *Utility theory for decision making*. New York: Wiley.

Han, J., Pei, J., Yin, Y., & Mao, R. (in press). Mining frequent patterns without candidate generation: A frequent-pattern tree approach. In *Data Mining and Knowledge Discovery*. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Haubl, G., & Murray, K. (2001). Recommending or persuading? The impact of a shopping agent's algorithm on user behavior. In *Proceedings of the 3rd ACM Conference on Electronic Commerce* (pp. 163–170). New York: ACM.

Joachims, T., Freitag, D., & Mitchell, T. (1997). Webwatcher: A tour guide for the World Wide Web. In *Proceedings of the 15th International Joint Conference On Artificial Intelligence* (pp. 770–775). San Francisco: Morgan Kaufmann.

Kramer, J., Noronha, S., & Vergo, J. (2000). A user-centered design approach to personalization. *Communications of the ACM, 43*(8), 45–48.

Lieberman, H. (1997). Autonomous interface agents. In *Proceedings of the ACM Conference on Computers and Human Interfaces (CHI-97)* (pp. 67–74). New York: ACM.

Luce, R. (1977). Conjoint measurement: A brief survey. In D. Bell, R. Keeney, & H. Raiffa (Eds.), *Conflicting Objectives in Decisions*. New York: Wiley.

Melville, P., Mooney, R., & Nagarajan, R. (2002). Content-boosted collaborative filtering for improved recommendations. In *Proceedings of the Eighteen National Conference on Artificial Intelligence* (pp. 187–192). San Francisco: Morgan Kaufmann.

Mobasher, B., Dai, H., Luo, T., & Nakagawa, M. (2002). Discovery and evaluation of aggregate usage profiles for Web personalization. In *Data Mining and Knowledge Discovery, 6*(1), 61–82.

Perkowitz, M., & Etzioni, O. (2000). Towards adaptive Web sites: Conceptual framework and case study. *Artificial Intelligence, 118*(1–2), 245–275.

Personalization Consortium (n.d.). Retrieved from http://www.personalization.org/

Pirolli, P., Pitkow, J., & Rao, R. (1996). Silk from a sow's ear: Extracting useful structures from the Web. In *Proceedings of Human Factors in Computing Systems (CHI-96)* (pp. 118–125). New York: ACM.

Saaty, T. (1980). *The analytic hierarchy process*. New York: McGraw–Hill.

Schafer, J., Konstan, J., & Riedl, J. (2001). E-commerce recommendation applications. *Data Mining and Knowledge Discovery, 5*, 115–153.

Spiliopoulou, M., & Faulstich, L. (1999). WUM: A tool for Web utilization analysis. In *Proceedings of the EDBT Workshop on the Web and Databases* (pp. 184–203). New York: Springer.

Weiss, G. (1999). *Multiagent systems: A modern approach to distributed artificial intelligence*. Cambridge, MA: MIT Press.

Westin, A. (1967). *Privacy and freedom*. New York: Atheneum Press.

Physical Security

Mark Michael, *King's College*

INTRODUCTION

Digital information is at the heart of every Internet transaction. The confidentiality, integrity, and availability of that information depends on the security of the following physical constituents of any computing environment:

1. hardware, in the broadest sense—machines, storage media, and transmission media;
2. the physical expression of the zeroes and ones that represent digital information (data and programs);
3. electricity, without which no digital information could change, move, or incite action; and
4. humans and the information they possess to run the system.

Internet security can be divided into two distinct areas: cybersecurity and physical security. The former term pertains to threats and defenses mounted via the same channels as legitimate exchanges of digital information. Encryption of information falls into this category. The role of physical security is to guard the four physical ingredients just outlined in two ways. First, it must protect the integrity and availability of resources for legitimate use. Second, it must prevent the misuse of resources, for example, by breaches of confidentiality or theft of services.

Physical security and cybersecurity complement one another. Where an organization's control over the physical ingredients ends, encryption and the like must take over. When cyberdefenses are strengthened, physical vulnerabilities become more inviting targets. Physical security serves cybersecurity. A breach of physical security, such as a password in the trash, can give a cyberattacker a foothold. The advent of biometrics and smart cards can be viewed either as an expansion of physical security into cybersecurity territory or as a blurring of the line between the two forms of security.

Physical security issues extend as far as an organization's resources. Because human knowledge is one of those assets, physical security concerns can span as far as information can spread. For instance, sensitive information could be revealed by an indiscreet question posted on a newsgroup. Thus, physical security is not constrained by a geographical footprint.

Physical security is intractable in the sense that certain events cannot be prevented. We cannot stop someone from demanding an off-duty employee's password at gunpoint, for instance. Redundancy is the last line of defense for the integrity and availability of resources. Confidentiality, on the other hand, cannot be "backed up"; some damage, such the as revelation of personal information, can never be repaired.

PHYSICAL THREATS TO INTEGRITY AND AVAILABILITY OF RESOURCES

Basic Environmental Threats to Computing Resources

Four basic threats to the physical health of computing equipment are inappropriate temperature, inappropriate humidity, foreign particles, and water.

Table 1 Temperature Thresholds for Damage to Computing Resources

COMPONENT OR MEDIUM	SUSTAINED AMBIENT TEMPERATURE AT WHICH DAMAGE MAY BEGIN
Flexible disks, magnetic tapes, etc.	38°C (100°F)
Optical media	49°C (120°F)
Hard-disk media	66°C (150°F)
Computer equipment	79°C (175°F)
Thermoplastic insulation on wires carrying hazardous voltage	125°C (257°F)
Paper products	177°C (350°F)

Source: Data taken from National Fire Protection Association (1999).

Temperature and Humidity

The internal temperature of equipment can be significantly higher than that of the room air. Although increasing densities have brought decreasing currents at the integrated circuit level, dissipation of heat is still a major concern. If a cooling system fails, a vent is blocked, or moving parts create abnormal friction, temperature levels can rise rapidly.

Excessively high temperatures can decrease performance or even cause permanent damage to computer equipment and media. The severity of the damage increases with temperature and exposure time, and its onset depends on the type of resource, as detailed in Table 1. Media may be reconditioned to recover data, but the success rate drops rapidly above these thresholds. Magnetism—the essence of much data storage—can be affected by temperatures higher than those listed; therefore, damage to magnetic media occurs first in the carrier and binding materials. On the other hand, silicon—the foundation of current integrated circuitry—will lose its semiconductor properties at significantly lower temperatures than what it takes to melt the solder that connects a chip to the rest of the computer.

To put these temperatures in perspective, some heat-activated fire suppression systems are triggered by ambient temperatures (at the sensor) as high as 71°C (160°F). Even in temperate climates, the passenger compartment of a sealed automobile baking in sunlight can reach temperatures in excess of 60°C (140°F). If media or a mobile computer is directly in sunlight and absorbing radiant energy, the heating is more rapid and pronounced, especially if the encasing material is a dark color, which, in the shade, would help radiate heat. (Direct sunlight is bad for optical media even at safe temperatures.)

Although excessive heat is the more common culprit, computing equipment also has a minimum temperature for operation. Frigid temperatures can permanently damage mobile components (e.g., the rechargeable battery of a laptop computer), even when (in fact, *especially* when) they are not in use. Plastics can also become more brittle and subject to cracking with little or no impact.

High humidity threatens resources in different ways. For electrical equipment, the most common problem is the long-term corrosive effect. If condensation forms, however, it brings the dangers posed by water (detailed later). Magnetic media deteriorate by *hydrolysis,* in which polymers "consume" water; the binder ceases to bind magnetic particles to the carrier and sheds a sticky material (which is particularly bad for tapes). Obviously, the rate of decay increases with humidity (and, as for any chemical process, temperature). Formation of mold and mildew can damage paper-based records, furniture, and so on. It can also obstruct reading from optical media. A bigger concern for optical media is corrosion of the metallic reflective layer. In tropical regions, there are even documented cases of fungi burrowing in CDs and corrupting data; high humidity promotes the fungal growth.

On the other hand, very low humidity may change the shape of some materials, thereby affecting performance. A more serious concern is that static electricity is more likely to build up in a dry atmosphere.

Foreign Particles

Foreign particles, in the broad sense intended here, range from insects down to molecules that are not native to the atmosphere. The most prevalent threat is dust. Even fibers from fabric and paper are abrasive and slightly conductive. Worse are finer, granular dirt particles. Manufacturing by-products, especially metal particles with jagged shapes, are worse yet. A residue of dust can interfere with the process of reading from media. Dirty magnetic tape can actually stick and break. Rotating media can be ground repeatedly by a single particle; a head crash is a possible outcome. A massive influx of dust (such as occurred near the World Trade Center) or volcanic ash can overwhelm the air-filtering capability of *HVAC* (heating, ventilation, and air-conditioning) systems.

Dust surges that originate within a facility due to construction or maintenance work are not only more likely than nearby catastrophes, they can also be more difficult to deal with because there is no air filter between the source and the endangered equipment. A common problem occurs when the panels of a suspended ceiling are lifted and particles rain down.

Keyboards are convenient input devices—for dust and worse. The temptation to eat or drink while typing only grows as people increasingly multitask. Food crumbs are stickier and more difficult to remove than ordinary dust. Carbonated drinks are not only sticky but also far more corrosive than water. In industrial contexts, other handborne substances may also enter.

Some airborne particles are liquid droplets or *aerosols*. Those produced by industrial processes may be highly corrosive. A more common and particularly pernicious aerosol is grease particles from cooking, perhaps in an employee lunchroom; the resulting residue may be less obvious than dust and cling more tenaciously.

Smoke consists of gases, particulates, and possibly aerosols resulting from *combustion* (rapid oxidation, usually accompanied by glow or flame) or *pyrolysis* (heat-induced physiochemical transformation of material, often prior to combustion). The components of smoke, including that from tobacco products, pose all the hazards of dust and may be corrosive as well.

Removable storage media often leave the protection of a controlled environment. They can suffer from contact with solvents or other chemicals.

There is an ever-growing list of potential chemical, biological, and radiological contaminants, each posing its own set of dangers to humans. Most are eventually involved in storage or transportation mishaps. More and more are intentionally used in a destructive fashion. Even if humans are the only component of the computing environment that is threatened, normal operations at a facility must cease until any life- or health-threatening contamination is removed.

Water

Water is a well-known threat to most objects of human design. Damage to paper products and the like is immediate. Mold and mildew will begin growing on certain damp materials. Sooner or later, most metals corrode (sooner if other substances, such as combustion by-products, are present).

The most critical problem is in energized electrical equipment. Water's conductive nature can cause a *short circuit* (a current that flows outside the intended path). When the improper route cannot handle the current, the result is heat, which will be intense if there is *arcing* (a luminous discharge from an electric current bridging a gap between objects). This may melt or damage items, even spawn an electrical fire.

Invasive water comes from two directions: rising from below and falling from above. Either may be the result of nature or human action. Floodwater brings two additional threats: its force and what it carries. The force of moving water and debris can do structural damage directly or indirectly, by eroding foundations. In some cases, natural gas lines are broken, which feed electrical fires started by short-circuiting. Most flood damage, however, comes from the water's suspended load. Whereas falling water, say from a water sprinkler or a leaking roof, is fairly pure and relatively easy to clean up, floodwater is almost always muddy. Fine particles (clays) cling tenaciously, making cleanup a nightmare. A dangerous biological component may be present if sewage removal or treatment systems back up or overflow or if initially safe water is not drained promptly. Another hazard is chemicals that may have escaped containment far upstream. When flooding or subsequent fire has disabled HVAC systems in the winter, ice formation has sometimes added further complications. Freezing water wedges items apart. Obviously, recovery is further delayed by the need to first thaw the ice.

Fire

Throughout history, fire has been one of the most important threats to human life, property, and activity when measured in terms of frequency, potential magnitude, and rapidity of spread. Fire presents a bundle of the previously mentioned environmental threats. By definition, combustion involves chemical and physical changes in matter, in other words, destruction of what was. Even away from the site of actual combustion, heat can do damage, as detailed earlier. Smoke can damage objects far from the site of combustion. More critical to humans are the irritant, toxic, asphyxial, and carcinogenic properties of smoke; it is the leading cause of death related to fire. With the advent of modern synthetic materials, fires can now produce deadlier toxins. Hydrogen cyanide, for instance, is approximately 25 times more toxic than carbon monoxide.

Sometimes the cure can be worse than the disease. If water is the suppressing agent, it can wreak havoc on adjacent rooms or lower floors that suffered no fire damage at all. Some modern fire suppressants decompose into dangerous substances. A comprehensive tome on fire is Cote (1997).

Power Anomalies

Electrical power is to electrical equipment what oxygen is to humans. Both the quantity and quality of electricity supplied to equipment are important. Just as humans can suffer, even die, from too much or too little air pressure, electrical equipment may malfunction or be permanently damaged when fed the wrong amount of current or voltage. This accounts for approximately half of computer data loss. Just as a properly pressurized atmosphere may carry constituents harmful to the immediate or long-term health of people, problems can arise when the power being supplied to a computer is itself conveying "information" in conflict with the digital information of interest.

Power Fluctuations and Interruptions

Low-voltage equipment such as telephones, modems, and networks are susceptible to small changes in voltage. Integrated circuits operate on very low currents (measured in milliamps); they can be damaged by minute changes in current. Power fluctuations can have a cumulative effect on circuitry over time, termed "electronic rust." Of the data losses due to power fluctuations, about three fourths of culpable events are drops in power.

The power grid, even under normal conditions, will deliver transients created as part of the continual balancing act performed in distributing power. Loose connections, wind, tree limbs, and errant drivers are among causes of abnormalities. Both the power grid and communications can be affected by so-called space weather. The Earth's magnetic field captures high-energy particles from the solar wind, shielding most of the planet while focusing it near the magnetic poles. Communications satellites passing between oppositely charged "sheets" of particles (seen as the Aurorae Borealis and Australis) may suffer induced currents, even arcing; one was permanently disabled in

1997. A *surge* (sudden increase in current) due to a 1989 geomagnetic storm blew a transformer, which in turn brought down the entire HydroQuébec electric grid in 90 seconds. The periods of most intense solar activity generally coincide with *Solar Max,* when the cycle of sunspot activity peaks every 10.8 years (on the average). The most recent peak was in July 2000.

A more frequent source of surges is lightning. In addition to direct hits on power lines or a building, near-misses can travel through the ground and enter a building via pipes, telecommunication lines, or nails in walls. Even cloud-to-cloud bolts can induce voltage on power lines.

Although external sources are the obvious culprits, the reality is that most power fluctuations originate within a facility. A common circumstance is when a device that draws a large inductive load is turned off or on; thermostatically controlled devices, such as fans and compressors for cooling equipment, may turn off and on frequently.

An *ESD* (electrostatic discharge) of *triboelectricity* (static electricity) generated by friction can produce electromagnetic interference (see below) or a *spike* (momentary increase in voltage) of surprisingly high voltage. Among factors contributing to a static-prone environment are low relative humidity (possibly a consequence of heating) and synthetic fibers in floor coverings, upholstery, and clothing. Especially at risk is integrated circuitry that has been removed from its antistatic packaging just before installation.

Electromagnetic Interference

Digital and analog information is transmitted over conductive media by modulating an electrical current or is broadcast by modulating an electromagnetic wave. Even information intended to remain within one device, however, may become interference for another device. All energized wires have the potential to broadcast, and all wires, energized or not, may receive signals. The messages may have no more meaning than the "snow" on a television screen. Even with millions of cell phones on the loose, much of the "electromagnetic smog" is incidental, produced by devices not designed to broadcast information.

The terms *EMI* (electromagnetic interference) and *RFI* (radio frequency interference) are used somewhat interchangeably. *Electrical noise* usually indicates interference introduced via the power input, though radiated energy may have been among the original sources of the noise; this term is also used with regard to small spikes. *EMC* (electromagnetic compatibility) is a measure of a component's ability neither to radiate electromagnetic energy nor to be adversely affected by electromagnetic energy originating externally. Good EMC makes for good neighbors. The simplest example of incompatibility is *crosstalk,* when information from one cable is picked up by another cable. By its nature, a digital signal is more likely to be received noise-free than an analog signal.

EMI from natural sources is typically insignificant (background radiation) or sporadic (like the pop of distant lightning heard on an amplitude modulated radio). Occasionally, solar flares can muddle or even jam radio communications on a planetary scale, especially at Solar Max. Fortunately, a 12-hour window for such a disruption can be predicted days in advance.

Most EMI results from electrical devices or the wires between. Power supply lines can also be modulated to synchronize wall clocks within a facility; this information can interfere with the proper functioning of computer systems. For radiated interference, mobile phones and other devices designed to transmit signals are a major hazard; according to Garfinkel (2002), they have triggered explosive charges in fire-extinguisher systems. Major high-voltage power lines generate fields so powerful that their potential impact on human health has been called into question. Motors are infamous sources of conducted noise, although they can radiate interference as well. For an introduction to electromagnetic interference, see the glossary and the chapter "EMI Shielding Theory" in Chomerics (2000).

Computing Infrastructure Problems

Hardware failures will still occur unexpectedly despite the best efforts to control the computing environment. Hard-drive crashes are one of the most infamous malfunctions, but any electronic or mechanical device in the computing environment can fail. In this regard, critical support equipment, such as HVAC, must not be overlooked. After the attack on the Pentagon Building, continued computer operations hinged on stopping the hemorrhage of chilled water for climate control.

The Internet exists to connect computing resources. Loss of telecommunications capabilities effectively nullifies any facility whose sole purpose is to serve the outside world. The difficulty may originate internally or externally. In the latter case, an organization must depend on the problem-solving efficiency of another company. In situations in which voice and data are carried by two separate systems, each is a possible point of failure. Although continuity of data transfer is the highest priority, maintenance of voice communications is still necessary to support the computing environment.

Physical Damage

Computers can easily be victims of premeditated, impulsive, or accidental damage. The list of possible human acts ranges from removing one key on a keyboard to formatting a hard drive to burning down a building. The focus here is on the fundamental forces that can damage equipment. Although computers and their components have improved considerably in shock resistance, there are still many points of potential failure due to shock. Hard drives and laptop *LCD* (liquid crystal display) screens remain particularly susceptible. More insidious are protracted, chronic vibrations. These can occur if fixed equipment must be located near machinery, such as HVAC equipment or a printer. Mobile equipment that is frequently in transit is also at higher risk. Persistent vibrations can loosen things, notably screws, that would not be dislodged by a sharp blow.

Removable storage media are more vulnerable to damage because they are more mobile and delicate. They can be damaged by bending, even if they appear to return to their original shape. Optical media, for instance, can

suffer microscopic cracking or *delamination* (separation of layers). Scratches and cracks on the data ("bottom") side of the disc will interfere with reading data. Cracks or delamination may also allow the incursion of air and the subsequent deterioration of the reflective layer. That layer is actually much closer to the label ("top") side and therefore can be easily damaged by scratches or inappropriate chemicals (from adhesives or markers) on the label side.

Although physical shocks can affect magnetic media by partially rearranging ferromagnetic particles, a far more common cause for magnetic realignment is, of course, magnetic fields. The Earth's magnetic field, averaging about 0.5 Gauss at the surface, does no long-term, cumulative damage to magnetic media. Certain electrical devices pose hazards to magnetic media; among these are electromagnets, motors, transformers, magnetic imaging devices, metal detectors, and devices for activating or deactivating inventory surveillance tags. (X-ray scanners and inventory surveillance antennae do not pose a threat.) Degaussers (bulk erasers) can produce fields in excess of 4,000 Gauss, strong enough to affect media not intended for erasure. Although magnetic media are the obvious victims of magnetic fields, some equipment can also be damaged by strong magnetic fields.

Local Hazards

Every location presents a unique set of security challenges. There are innumerable hazards the probability and impact of which are location-dependant. Often, a pipeline, rail line, or road in the immediate vicinity carries the most likely and most devastating potential hazard. Two of the local hazards with the greatest impact on human life, property, and activity are flooding and geological events.

Flooding

As many have learned too late, much flood damage occurs in areas not considered flood-prone. Government maps depicting flood potential are not necessarily useful in assessing risk, because they can quickly become outdated. One reason is construction in areas with no recorded flood history. Another is that urbanization itself changes drainage patterns and reduces natural absorption of water.

Small streams react first and most rapidly to rainfall or snowmelt. Even a very localized rain event can have a profound effect on an unnoticed creek. Perhaps the most dangerous situation is in arid regions, where an intermittent stream may be dry or nearly dry on the surface for much of the year. A year's worth of rain may arrive in an hour. Because such flash floods may come decades apart, the threat may be unrecognized or cost-prohibitive to address.

Usually, advance warning of floods along large rivers is better than for the small rivers that feed them. Having a larger watershed, large rivers react more slowly to excessive rain or rapidly melting snow. Formation of ice jams, breaking of ice jams, structural failure of dams, and landslides or avalanches into lakes, however, can cause a sudden, unexpected rise in the level of a sizeable river.

Coastal areas are occasionally subjected to two other types of flooding. The storm surge associated with a hurricane-like storm (in any season) can produce profound and widespread damage, but advanced warning is usually good enough to make appropriate preparations. Moving at 725 km (450 miles) per hour on the open ocean, *tsunamis* (seismic sea waves) caused by undersea earthquakes or landslides arrive with little to no warning and can be higher than storm surges. Although tsunamis most often strike Pacific coastlines, a much larger (and rarer) *mega-tsunami* could effect much of the Atlantic if a volcano in the Canary Islands collapses all at once.

An urban area is at the mercy of an artificial drainage system, the maintenance of which is often at the mercy of a municipality. A violent storm can itself create enough debris to greatly diminish the system's drainage capacity.

Not all flooding originates in bodies of water. Breaks in water mains can occur at any time, but especially during winter freeze-thaw cycles or excavation. Fire hydrants can be damaged by vehicles. Pipes can leak or commodes overflow. Although safest from rising water, the top floor is the first affected if the roof leaks, collapses, or is blown away.

Geological Events

Geological hazards fall into a number of categories. These events are far more unpredictable than meteorological events, although some, notably landslides and mudslides, may be triggered by weather. Earthquakes can have widespread effects on infrastructure. The damage to an individual structure may depend more on *where* it was built than on *how*. Buildings on fill dirt are at greater risk because of potential *liquefaction,* in which the ground behaves like a liquid. Earthquake predictions are currently vague as to time and location.

Landslides and mudslides are more common after earthquakes and rainstorms, but they can occur with no obvious triggering event. Anticipating where slides might occur may require professional geological consultation. As an illustration, a cliff with layers of clay dipping toward the face of the cliff is an accident waiting to happen.

Volcanic ash is one of the most abrasive substances in nature. It can occasionally be carried great distances and in great quantities. If it does not thoroughly clog up HVAC air filters between outside and inside air domains, it may still be tracked in by people. Most volcanic eruptions are now predictable.

Humans

Humans are often referred to as the "weakest link" in computing security, for they are the computing environment component most likely to fail. Despite their flaws, humans have always been recognized as an essential resource. Before the attacks on New York and Washington, however, the sudden disappearance of large numbers of personnel was simply not anticipated by most business continuity planners or disaster recovery planners. All planners, whether focused on preservation of processes or assets, now have a different outlook on preservation of life.

Aside from mass slaughter, there are other circumstances in which human resources may be lacking. Severe weather may preclude employees from getting to work. Labor disputes may result in strikes. These may be beyond the direct control of an organization if the problems

are with a vendor from whom equipment has been bought or leased or with a contractor to whom services have been outsourced. A different kind of discontinuity in human expertise can come with a change of vendors or contractors.

Even the temporary absence or decreased productivity of individuals soon adds up to a major business expense. Employers may be held responsible for a wide range of occupational safety issues. Those specific to the computing environment include

1. carpal tunnel syndrome (from repetitive actions, notably typing),
2. back and neck pain (from extended use of improper seating), and
3. eye strain and headaches (from staring at a computer screen for long periods).

PHYSICAL MEANS OF MISAPPROPRIATING RESOURCES

I now turn to the misappropriation of assets that can be possessed in some sense—physical objects, information, and computing power. (Some acts, such as physical theft, also impinge on availability). Misuse may entail use by the wrong people or by the right people in the wrong way. The transgressions may be without malice. A pilferer of "excess" computing power may view his or her actions as a "victimless crime." In other cases, insiders create new points of presence (and, therefore, new weak points) in an attempt to possess improved, legitimate access. See Skoudis (2002) for discussions of many of these issues.

Unauthorized Movement of Resources

For computing resources, theft comes in several forms. Outsiders may break or sneak into a facility. Insiders may aid a break-in, may break into an area or safe where (or when) they are not entitled to access, or they may abuse access privileges that are a normal part of their job. Physical objects may be removed. Information, whether digital or printed, may be duplicated or merely memorized; this is classified as theft by copying.

A different situation is when items containing recoverable data have been intentionally discarded or designated for recycling. The term *dumpster diving* conjures up images of an unauthorized person recovering items from trash bins outside a building (although perhaps still on an organization's property). In fact, discarded items can also be recovered from sites inside the facility by a malicious insider. At the other extreme, recovery could, in theory, take place thousands of miles from the point at which an object was initially discarded. A large fraction of the "recycled" components from industrialized countries actually end up in trash heaps in Third World countries. The legality of dumpster diving depends on local laws and on the circumstances under which an item was discarded and recovered.

Perhaps the most obvious candidate for theft is removable storage media. As the data density of removable storage media increases, so does the volume of information that can be stored on one item and, therefore, the ease with which a vast amount of information can be stolen. Likewise, downloading from fixed media to removable media can also be done on a larger scale, facilitating theft by copying.

By comparison, stealing hardware usually involves removing bigger, more obvious objects, such as computers and peripherals, with the outcome being more apparent to the victim. Garfinkel (2002) reports thefts of random access memory (RAM); if not all the RAM is removed from a machine, the loss in performance might not be noticed immediately.

Social Engineering and Information Mining

Human knowledge is an asset less tangible than data on a disk but worth possessing, especially if one is mounting a cyberattack. An attacker can employ a variety of creative ways to obtain information. *Social engineering* involves duping someone else to achieve one's own illegitimate end. The perpetrator—who may or may not be an outsider—typically impersonates an insider having some privileges ("I forgot my password . . . "). The request may be for privileged information ("Please remind me of my password . . . ") or for an action requiring greater privileges ("Please reset my password . . . "). Larger organizations are easier targets for outsiders because no one knows everyone in the firm. Less famous than social engineering are methods of mining public information. Some information must necessarily remain public, some should not be revealed, and some should be obfuscated.

Domain name service information related to an organization—domain names, *IP* (Internet protocol) addresses, and contact information for key information technology (IT) personnel—must be stored in an online "whois" database. If the name of a server is imprudently chosen, it may reveal the machine's maker, software, or role. Such information makes the IP addresses more useful for cyberattacks. Knowing the key IT personnel may make it easier to pose as an insider for social engineering purposes.

Currently, the most obvious place to look for public information is an organization's own Web site. Unless access is controlled so that only specific users can view specific pages, anyone might learn about corporate hardware, software, vendors, and clients. The organizational chart and other, subtler clues about corporate culture may also aid a social engineering attack. Of course, this information and more may be available in print.

Another dimension of the Internet in which one can snoop is newsgroup bulletin boards. By passively searching these public discussions ("lurking"), an attacker might infer which company is running which software on which hardware. He or she may instead fish actively for information. An even more active approach is to provide disinformation, leading someone to incorrectly configure a system.

Unauthorized Connections and Use

Wiretapping involves making physical contact with guided transmission media for the purposes of intercepting information. Wired media are relatively easy to tap, and

detection (other than visual inspection of all exposed wires) may be difficult. Contrary to some rumors, fiber-optic cable remains far more difficult to tap, and detection (without visual inspection) is highly likely; any light that can be made to "leak" from a cable is not useable for recovering data.

A specific type of wiretapping is a *keyboard monitor,* a small device interposed between a computer and its keyboard that records all work done via the keyboard. The attacker (or suspicious employer) must physically install the item and access it to retrieve stored data. (Hence, keyboard logging is more often accomplished by software.)

A variation on wiretapping is to use connectivity hardware already in place, such as a live, unused *LAN* (local area network) wall jack; a live, unused hub port; a LAN-connected computer that no longer has a regular user; and a computer in use but left unattended by the user currently logged on. For the perpetrator, these approaches involve varying degrees of difficulty and risk. The second approach may be particularly easy, safe, and reliable if the hub is in an unsecured closet, the connection is used for sniffing only, and no one has the patience to check the haystack for one interloping needle.

Phone lines are connectivity hardware that is often overlooked. A naïve employee might connect a modem to an office machine so it can be accessed (for legitimate reasons) from home. This gives outsiders a potential way around the corporate firewall. Even IT administrators who should know better leave "back-door" modems in place, sometimes with trivial or no password protection. Sometimes the phone service itself is a resource that is misappropriated. Although less common now, some types of *PBX* (private branch exchange) can be "hacked," allowing an attacker to obtain free long-distance service or to mount modem-based attacks from a "spoofed" phone number.

A final asset is an adjunct to the phone service. Employee voice mail, even personal voice mail at home, has been compromised for the purpose of obtaining sensitive information (e.g., reset passwords).

Appropriate access through appropriate channels does not imply appropriate use. One of the biggest productivity issues nowadays is employee e-mail and Internet surfing unrelated to work. If prohibited by company policy, this can be viewed as misappropriation of equipment, services, and, perhaps most important, *time*. Although text-based e-mail is a drop in the bucket, downloading music files can "steal" considerable bandwidth; this is especially a problem at those academic institutions where control of students' Internet usage is minimal.

Eavesdropping

Eavesdropping originally meant listening to something illicitly. Although capture of acoustic waves (perhaps with an infrared beam) is still a threat, the primary concern in the computing environment involves electronically capturing information without physical contact. Unguided transmission media such as microwave (whether terrestrial or satellite), radio (the easiest to intercept), and infrared (the hardest to intercept) should be considered fair game for outsiders to eavesdrop; such transmissions must be encrypted if security is a concern. Among guided transmission media, fiber-optic cable stands alone for its inability to radiate or induce any signal on which to eavesdrop. Therefore, the interesting side of eavesdropping is tempest emissions. Electrical devices and wires have long been known to emit electromagnetic radiation, which is considered "compromising" if it contains recoverable information. Mobile detectors have been used to locate radios and televisions (where licensing is required) or to determine the stations to which they are tuned. Video displays (including those of laptops) are notorious emitters; inexpensive equipment can easily capture scan lines, even from the video cable to an inactive screen.

The term *tempest* originated as the code word for a U.S. government program to prevent compromising emissions. (Governments are highly secretive in this area; contractors need security clearance to learn the specifications for equipment to be tempest-certified.) Related compromising phenomena are as follows:

1. *hijack*—signals conducted through wires (and perhaps the ground, as was noted during World War I);
2. *teapot*—emissions intentionally caused by an adversary (possibly by implanted software); and
3. *nonstop*—emissions accidentally induced by nearby radio frequency (RF) sources.

One attack is to irradiate a target to provoke resonant emissions—in other words, intentional nonstop. (This is analogous to how an infrared beam can expropriate acoustic information.) Interestingly, equipment certified against passive tempest eavesdropping is not necessarily immune to this more active attack. (Compare the infrared device to a parabolic microphone, which is merely a big ear.) Although these emissions were formerly the concern only of governments, increasingly less expensive and more sophisticated equipment is making corporate espionage a growing temptation and concern. An excellent introduction to this area is chapter 15 of Anderson (2001). A well-known portal for tempest information is McNamara (2002).

PREVENTIVE MEASURES

To expand George Santayana's famous quote, those who are ignorant of history are doomed to repeat it, but those who live in the past are also doomed. Although an understanding of past disasters is essential, not all that *will* happen (in your neighborhood or in the world) *has* happened. The key to preventing physical breaches of confidentiality, integrity, and availability of computing resources is to anticipate as many bad scenarios as possible. A common flaw is to overlook plausible combinations of problems, such as the incursion of water while backup power is needed.

History has taught us that, regardless of the time, effort, and money invested, preventing *all* bad events is impossible; there *will* be failures. For integrity and availability of resources, redundancy can be used as a parachute

when the worst-case scenario becomes reality. Unfortunately, there is no comparable preventive measure for confidentiality.

Control and Monitoring of Physical Access and Use

There are several philosophical approaches to physical access control, which can be used in combination with one another:

1. Physical contact with a resource is restricted by putting it in a locked cabinet, safe, or room; this would deter even vandalism.
2. Contact with a machine is allowed, but it is secured (perhaps permanently bolted) to an object difficult to move; this would deter theft. A variation of this allows movement, but a motion-sensored alarm sounds.
3. Contact with a machine is allowed, but a security device controls the power switch.
4. A machine can be turned on, but a security device controls log-on. Related to this is the idea of having a password-protected screensaver running while the user is away from the machine.
5. A resource is equipped with a tracking device so that a sensing portal can alert security personnel or trigger an automated barrier to prevent the object from being moved out of its proper security area.
6. An object, either a resource or a person, is equipped with a tracking device so that his, her, or its current position can be monitored continually.
7. Resources are merely checked in and out by employees, for example by scanning barcodes on items and ID cards, so administrators know at all times of who has what, but not necessarily where they have it.

Yet another approach can be applied to mobile computers, which are easier targets for theft. More and more high-density, removable storage options are available, including RAM-disks, DVD-RAMs, and memory sticks. This extreme portability of data can be turned to an advantage. The idea is to "sacrifice" hardware but preserve the confidentiality of information. If no remnant of the data is stored with or within a laptop (which may be difficult to ensure), the theft of the machine from a vehicle or room will not compromise the data. The downside is that the machine is removed as a locus of backup data.

There are also a multitude of "locks." Traditional locks use metal keys or require a "combination" to be dialed on a wheel or punched on an electronic keypad. Another traditional "key" is a photo ID card, inspected by security personnel. Newer systems require the insertion or proximity of a card or badge; the types of cards include magnetic stripe cards, memory cards, optically coded cards, and smart cards (either contact or contactless). The most promising direction for the future appears to be biometric devices, the subject of a separate article; a major advantage of these is that they depend on a physiological or behavioral characteristic, which cannot be forgotten or lost and is nearly impossible to forge.

To paraphrase General George C. Patton, any security device designed by humans can be defeated by humans. Each type of locking device has its own vulnerabilities and should be viewed as a *deterrent*. In some cases, even an inexpensive, old-fashioned lock is an adequate deterrent—and certainly better than nothing (as is often the case with wiring cabinets). In assessing a candidate for a security device or architecture, the time, resources, and sophistication of a likely, hypothetical attacker must be correlated with both the security scheme *and* the assets it protects.

An example may be helpful. To determine the suitability of smart cards, first research the many potential attacks on smart cards and readers. Then estimate how long an outsider or malicious insider might have unsupervised access to a smart card or reader of the type used or in actual use. Finally, make a guess as to whether the assets at stake would motivate an adversary to invest in the necessary equipment and expertise to perform a successful attack given the level of access they have.

It is sometimes appropriate for an organization to allow public access on some of its computers. Such computers should be on a separate LAN, isolated from sensitive resources. Furthermore, to avoid any liability issues, the public should not be afforded unrestricted access to the Internet.

A different aspect of access is unauthorized connections. A multipronged defense is needed. Checking for renegade modems can be done either by visually inspecting every computer or by war-dialing company extensions. Hubs must be secured and their ports should be checked to verify that they are used only by legitimate machines. Unused jacks or jacks for unused computers must be deactivated. Computers that are no longer on the LAN must be locked away or at least have their hard drives sanitized. To prevent wiretapping, all wires not in secured spaces should be enclosed in pipes (which can themselves be protected against tampering). Unprotected wires can periodically be tested by sending pulses down the wires; exhaustive visual inspections are impractical.

A more complex issue is that of improper use of services, especially e-mail and Internet access, whose proper use may be an essential part of work-related duties. Companies are within their rights to limit or track the usage of their resources in these ways, even if employees are not forewarned. Many employers monitor e-mail passing through company hardware, even that for an employee's personal e-mail account. In addition, they use *activity monitors,* software to record keystrokes, to capture screen displays, or to log network access or use of applications. (These monitoring activities can in turn be detected by employees with suitable software.) Alternatively, inbound or outbound Internet traffic can be selectively blocked, filtered, or *shaped*; the last is the least intrusive because it limits the portion of bandwidth that can be consumed by certain services while not prohibiting them entirely.

Control and Monitoring of Environmental Factors

HVAC systems should have independently controlled temperature and relative humidity settings. Each variable should be monitored by a system that can issue alerts

when problems arise. Ideally, HVAC units should be installed in pairs, with each unit being able to carry the load of the other should it malfunction.

Although some information is only of transitory value, other data, such as official records of births, deaths, marriages, and transfers of property ownership, should be kept in perpetuity. Standards for long-term preservation of data stored in magnetic or optical format are far stricter than guidelines for ordinary usage. As a sample, for preservation, the prescribed allowable temperature variation in 24 hours is a mere $\pm 1^{\circ}$C (2°F). See International Advisory Committee for the UNESCO Memory of the World Programme (2000) for detailed preservation guidelines. One such guideline is that magnetic media, both tapes and disks, be stored in an upright orientation (i.e., with their axes of rotation horizontal). The exclusion of light is important for extending the useful life of optical media incorporating dyes (writeable discs). All media should be stored in containers that will not chemically interact with the media. Projected life spans for properly archived media are considered to be 5–10 years for floppy diskettes, 10–30 years for magnetic tapes, and 20–30 years for optical media. These estimates are conservative to ensure creation of a new copy before degradation is sufficient to invert any bits.

For optical media, life expectancies are extrapolated from accelerated aging tests based on assumptions and end-of-life criteria that may be invalid. Numerous factors influence longevity. Write-once formats have greater life expectancies than rewriteable formats. The bit-encoding dye phthalocyanine (appearing gold or yellowish green) is less susceptible than cyanine (green or blue-green) to damage from light after data has been written; yet manufacturers' claimed life expectancies of up to 300 years are not universally accepted. What appears to be a major determiner of longevity is the original quality of the stored data. This in turn depends on the quality of the blank disc, the quality of the machine writing the data, and speed at which data was written. Hartke (2001) gives an enlightening look at the complexities of this issue.

All archived data of critical importance should be sampled periodically and backed up *well before* the rate of correctable errors indicates that data might be unrecoverable at the next sampling. Even physically perfect data has been effectively lost because it outlived the software or hardware needed to read it. Therefore, before its storage format becomes obsolete, the data must be converted to an actively supported format.

There are devices or consumable products for cleaning every type of storage medium and every part of a computer or peripheral device. Backup tapes that are frequently overwritten should be periodically removed from service to be tested on a *tape certifier,* which writes sample data to the tape and reads it back to detect any errors; some models incorporate selective cleaning as an option. Read-write heads for magnetic media typically need to be cleaned far more often than the medium that moves by them. For optical media, clean discs are usually the concern. Compressed air should not be used; the resulting drop in temperature produces a *thermal shock* (rapid temperature change) for the disc. If the problem is scratches rather than dirt, polishing may be required.

Keeping a computing area free of foreign particles is a multifaceted task. Air filters should remove fine dust particles because outdoor dust is brought in on clothes and shoes. Filters must be cleaned or replaced on a regular schedule. Periodically, air-heating equipment should be turned on briefly even when not needed. This is to incrementally burn off dust that would otherwise accumulate and be converted to an appreciable amount of smoke when the equipment is activated for the first time after a long period of disuse. Vacuuming of rooms and equipment should also involve filters. Food, drink, and tobacco products should be banned from the computing area.

Water detectors should be placed above and below a raised floor to monitor the rise of water. An automatic power shutdown should be triggered by a sensor that is lower than the lowest energized wire. Degaussers and any other equipment that produces strong magnetic fields should be kept in a room separate from any media not scheduled to be erased. Although the intensity of most magnetic fields decreases rapidly with distance, it is very difficult to shield against them. Likewise, computers should be kept away from sources of vibrations, including printers. If this cannot be arranged, vibration-absorbing mats can be placed under the computer or the offending device.

Health and Safety Issues

The humans in the computing environment have additional needs. Some general health issues that may arise are *sick building syndrome* (symptoms arising from toxic mold) and *Legionnaire's disease* (a form of pneumonia transmitted via mist and sometimes associated with large air conditioning systems). Human-friendly appointments pertinent to a computing environment include the following:

1. special keyboards or attachments that optimize wrist placement;
2. comfortable, adjustable chairs that properly support backs; and
3. special lighting, monitor hoods, or screen coverings that reduce glare and, therefore, eyestrain.

There is currently no consensus on the long-term effects of *extremely low-frequency* (ELF) emissions (below 300 Hz), magnetic fields emitted by a variety of devices, including high-tension lines and cathode ray tube monitors (but not LCD displays). Laboratory tests with animals have found that prolonged exposure to ELF fields may cause cancer or reproductive problems. Studies of pregnant CRT users have produced conflicting data. Pending conclusive evidence, some recommend keeping 60 centimeters (2 feet) away from such monitors, which may not be practical. There are similar concerns and uncertainty with regard to cellular phones. It is known that people with pacemakers should avoid devices creating strong magnetic fields, such as degaussers. Although the World Health Organization acknowledges the need for continued research in certain areas, its latest position is that there is no evidence of health risks associated with EMF exposures *below* the levels set forth by the

International Commission on Non-Ionizing Radiation Protection (1998).

Depending on the overall security architecture, the criticality of the facility, and the anticipated threats, it may be advisable to implement any or all of the following:

1. stationed or roving security guards;
2. surveillance cameras, monitored in real time and recorded on videotape;
3. motion detectors;
4. silent alarms (of the type used in banks); and
5. barriers that prevent unauthorized vehicles from approaching the facility.

Fire Preparedness

For the survival of people and inanimate objects, the most critical preparations are those regarding fire.

Fire Detection

Automatic fire detectors should be placed on the ceilings of rooms as well as in hidden spaces (e.g., below raised floors and above suspended ceilings). The number and positioning of detectors should take into account the location of critical items, the location of potential ignition sources, and the type of detector. Fire detectors are based on several technologies:

1. *Fixed-temperature heat detectors* are triggered at a specific temperature. Subtypes are
   (a) *fusible*—metal with a low melting temperature;
   (b) *line type*—insulation melts, completing a circuit; and
   (c) *bimetallic type*—bonding of two metals with unequal thermal expansion coefficients, bends when heated (the principle in metal-coil thermometers), completing a circuit (until cooled again).
2. *Rate-compensation detectors* trigger at a lower temperature if the temperature rise is faster.
3. *Rate-of-rise detectors* react to a rapid temperature rise, typically 7–8°C (12–15°F) per minute.
4. *Electronic spot type thermal detectors* use electronic circuitry to respond to a temperature rise.
5. *Flame detectors* "see" radiant energy. They are good in high-hazard areas. Subtypes are
   (a) *infrared*—can be fooled by sunlight, but less affected by smoke than ultraviolet detectors; and
   (b) *ultraviolet*—detects radiation in the 1850–2450 angstrom range (i.e., almost all fires).
6. Smoke detectors usually detect fires more rapidly than heat detectors. Subtypes are
   (a) *ionizing*—uses a small radioactive source (common in residences); and
   (b) *photoelectric*—detects obscuring or scattering of a light beam.

A third type of smoke detector is the *air-sampling type*. One version, the *cloud chamber smoke detector,* detects the formation of droplets around particles in a high-humidity chamber. Another version, the *continuous air-sampling smoke detector,* is particularly appropriate for computing facilities. It can detect very low smoke concentrations and report different alarm levels.

For high-hazard areas, there are also automatic devices for detecting the presence of combustible vapors or abnormal operating conditions likely to produce fire; said another way, they sound an alarm *before* a fire starts.

Some fire detectors, especially the fusible type, are integrated into an automatic fire suppression system. This means that the first alarm could be the actual release of an extinguishing agent. Because an event triggering a fire may also disrupt the electrical supply, fire detectors must be able to function during a power outage. Many fire detectors are powered by small batteries, which should be replaced on a regular schedule. Some components of detectors, such as the radioisotope in an ionizing smoke detector, have a finite life span; the viability of such a detector cannot be determined by pushing the "test" button, the purpose of which is merely to verify the health of the battery. Such detectors must be replaced according to the manufacturer's schedule.

Fire Prevention and Mitigation

Better than detecting a fire is preventing it from starting. The two things to avoid are high temperatures and low ignition points. It is usually possible to exclude highly flammable materials from the computing environment. Overheating is a possibility in almost any electrical device. In some cases a cooling system has failed or has been handicapped. In other cases, a defective component generates abnormal friction. The biggest threat comes from short circuits; the resulting resistance may create a small electric heater or incite arcing.

Some factors that may lead to a fire, such as short circuits within a machine or a wall, are beyond our control. Yet many precautions can be taken to lessen the chances of a fire. Vents should be kept unobstructed and air filters clean. Power circuits should not be asked to carry loads in excess of their rated capacity. Whenever possible, wires should run below a raised floor rather than on top of it. If wires must lie on a floor where they could be stepped on, a sturdy protective cover must be installed. In any case, wires should be protected from fatiguing or fraying. See National Fire Protection Association (1999) for fire prevention guidelines for the computing environment. As of this writing, the newest electrical code pertaining specifically to computing equipment is from the International Electrotechnical Commission (2001).

Many fires are actually the culmination of a protracted process. Another preventive measure is for employees to use their eyes, ears, noses, and brains. Damage to a power cord can be observed if potential trouble spots are checked. Uncharacteristic noises from a component may be symptomatic of a malfunction. The odor of baking thermoplastic insulation is a sign that things are heating up.

Given that a fire may have an external or deliberate origin, preventing the spread of fire is arguably more important than preventing its ignition. It certainly requires greater planning and expense. The key ideas are to erect fire-resistant barriers and to limit fuel for the fire between the barriers.

Table 2 Comparison of Types of Surge Protectors

TYPE OF SURGE PROTECTOR	CHARACTERISTICS
MOV (Metaloxide Varistor)	Inexpensive, easy to use, but progressively degrades from even minor surges (possibly leading to a fiery demise)
Gas Tube	Reacts quickly, can handle big surges, but may not deactivate until an alternating circuit polarity flip (which may mean the computer shuts down in the meantime)
SAD (Silicon Avalanche Diode)	Faster than an MOV (1 ns vs. 5 ns), but has a limited power capacity
Reactive Circuit	Also smoothes out noise but can only handle *normal-mode* surges (between hot and neutral lines) and may actually cause a *common-mode* surge (between neutral and ground lines), which is thought to be the more dangerous type of surge for desktop computers

For computing environments, the choice of construction materials, design, and techniques for mitigating the spread of fire should exceed the minimum standards dictated by local building codes. Because fires can spread through unseen open spaces, including ventilation systems, a *computing area* is defined to be all spaces served by the same HVAC system as a computing room. Air ducts within that system should have smoke dampers. The computing area must be isolated in a separate *fire division*. This means the walls must extend from the structural floor to the structural ceiling of the computer area and have a *one-hour rating* (resistance to an external fire for one hour). Care should be taken to ensure that openings where pipe and cables pass through the fire-resistant boundaries of the separate fire division are sealed with material that is equally fire-resistant.

Many fires affecting a computer area do not actually originate in that area. Even if a fire does not technically spread into a computing area, its products—heat, smoke, and *soot* (carbon deposits)—may. Consequently, the level of fire protection beyond the computing area is still of critical concern. *Fully sprinklered* buildings (protected by sprinkler systems throughout) are recommended. Concern should extend beyond the building if it is located in an area with high hazards, such as chemical storage or periodically dry vegetation. In the latter case, a *fire break* should be created around the building by removal of any vegatation likely to fuel a fire.

The standards prescribed by the National Fire Protection Association (1999) for fire protection of computing equipment set specifications for wall coverings, carpet, and furnishings (which are relaxed in fully sprinklered buildings). They also limit what other materials can be present. They do not take into account that even high-hazard areas have computers present. In interpreting those standards, determine which dangerous materials are absolutely essential for operations, and work to minimize any unnecessary hazards. Due to their potential contribution to fire (as well as being a more likely starting point for a fire), materials that could contribute to a *Class B* fire (including solvents, paints, etc.) should not be stored in a computing area except in a fireproof enclosure. Materials that could contribute to a *Class A* fire, such as paper, should be kept to the minimum necessary.

Raised floors are standard features of many computer facilities, allowing for cables to connect equipment without the need to cover cables to prevent fraying and electrical shorting. The use of junction boxes below the floor should be minimized, however. The needed equipment for lifting the heavy removable panels to gain access to the space between the raised floor and the structural floor must be easy to locate, even in the event of a fire.

Power Maintenance and Conditioning

The most basic necessity for the functioning of computer resources is maintenance of power. *Power conditioning* refers to smoothing out the irregularities of that power.

Surge Protectors and Line Filters

A *surge protector* is designed to protect against sudden increases in current. It forms a second line of defense, the circuit breaker being the first. Neither should be counted on to protect against a direct hit by lightning. There is no substitute for unplugging home computers during an electrical storm. A large building should have a separate lightning protection system in any case. Surge protectors are currently based on four technologies, described in Table 2.

Metaloxide varistor (MOV), gas tube, and silicon avalanche diode (SAD) surge protectors short out the surge and isolate it from the protected equipment. The reactive circuit type uses a large inductance to spread a surge out over time. All should have lights to indicate if they are in functioning order. MOVs and SADs are the types preferred for computing environments because of their reaction times. All surge protectors require a properly grounded electrical system in order to do their job.

Line filters clean power at a finer level, removing electrical noise entering through the line power. Their concern

is not extreme peaks and valleys in the alternating current (AC) sine wave, but modulation of that wave. Their goal is to restore the optimal sine shape. Power purity can also be fostered by adding circuits rather than filters. The most important precaution is to keep large machinery off any circuit powering computing equipment. If possible, it is preferable to have each computer on a separate circuit.

The dangers of static electricity can be reduced by inhibiting its buildup, providing ways for it to dissipate gradually (rather than discharge suddenly), or insulating vulnerable items. Antistatic techniques include the following:

1. keeping the relative humidity from dropping too low (below 40%);
2. avoiding the use of carpets and upholstery with synthetic fibers, or spraying them with antistatic sprays;
3. using antistatic tiles or carpets on floors;
4. not wearing synthetic clothing and shoes with soles prone to generating charges;
5. using an *ionizer* (which sends both positive and negative ions into the air as a neutralizing influence); and
6. keeping computers away from metal surfaces or covering metal surfaces with dissipative mats or coverings.

When installing electronic circuitry, technicians should ground themselves. A variety of conductive "garments" can be worn, including bracelets and straps for wrists and ankles, gloves, finger cots, and smocks.

Uninterruptible Power Supplies (UPS)

Although an *uninterruptible power supply,* by definition, counteracts a loss of power, it typically provides surge protection as well. This is accomplished by means of separate input and output circuits. The input circuit induces current in the output circuit. A UPS may also incorporate noise filtering. UPS systems fall into three categories. An *online* system separates the input and output with a buffer, a battery that is constantly in use and (almost) constantly being charged. This is analogous to a water tank providing consistent water pressure, regardless of whether water is being added to it. This is the original and most reliable design for a UPS. In the strictest sense, this is the only truly uninterruptible power supply; its transfer time (defined below) is zero milliseconds. An *offline* system sends the primary current straight through in normal circumstances, but transfers to backup power if its detection circuit recognizes a problem with the primary power. The problem might be a complete drop in primary power, but it might also be a spike, a surge, a *sag* (drop in voltage), or electrical noise. A *line interactive* system is similar to an offline system, but its output waveform will be a sine wave (as is the input waveform) rather than a square or step wave. Aside from its basic type, the most important characteristics of a UPS are its

1. *capacity*—how much of a load it can support (measured in volt-amps or watts);
2. *voltage*—the electromotive force with which the current is flowing (measured in volts);
3. *efficiency*—the ratio of output current to input current (expressed as a percentage);
4. *backup time*—the duration during which it can provide peak current (a few minutes to several hours);
5. *transfer time*—the time from the drop in primary power until the battery takes over (measured in milliseconds);
6. *battery life span*—how long it is rated to perform as advertised;
7. battery type—a small Ni-MH (nickel metal hydride) battery support of an individual machine, whereas lead-acid batteries for an entire facility may require a room of their own; and
8. output waveform—sine, square, or step (also known as a modified sine) wave.

A final consideration is the intended load: *resistive* (as a lamp), *capacitive* (as a computer), or *inductive* (as a motor). Because of the high starting current of an inductive load, the components of an offline UPS (with its square or step wave output) would be severely damaged. Actually, an inductive load will still have a similar but less severe effect on other types of UPS systems (with sine wave output).

Large battery systems may generate hydrogen gas, pose a fire hazard, or leak acid. Even a sealed, maintenance-free battery must be used correctly. It should never be fully discharged, it should always being recharged immediately after usage, and it should be tested periodically.

Some UPS systems feature scalability, redundancy, and interface software, which can

1. indicate the present condition of the battery and the main power source;
2. alert users when backup power is in operation, so that they can shut down normally; or
3. actually initiate a controlled shutdown of equipment prior to exhaustion of backup power.

A UPS should come with a warranty for equipment connected to the UPS; the value of any lost data is typically not covered.

When limited resources do now allow for all equipment to be on a UPS, the process of deciding which equipment is most critical and therefore most deserving of guaranteed power continuity should consider two questions. First, if power is lost, will appropriate personnel still receive automated notification of this event? Second, is the continued functioning of one piece of equipment moot if another component loses power?

The existence of any UPS becomes moot whenever someone accidentally flips the wrong switch. The low-cost, low-tech deterrent is switch covers, available in stock and custom sizes.

There are occasions (e.g., fires and floods) when power must be cut to all equipment except emergency lighting and fire detection and suppression systems (which should have self-contained power sources). This includes disconnecting a UPS from its load. Any intentional disruption of

power should be coordinated with computers via software to allow them to power down gracefully.

Electromagnetic Shielding

Because of their inherent vulnerability to interception, wireless transmissions should be encrypted (or scrambled, in the case of analog voice communication) if confidentiality, integrity, or authentication is essential. Electromagnetic shielding is in direct opposition to wireless communication. The purpose of shielding is to block outbound compromising emissions and inbound radiated interference. The key idea is a *Faraday cage* (i.e., a conductive enclosure). This can be accomplished at several levels.

Shielding entire rooms and buildings with metal, conductive wall coverings, conductive windows, and so forth to control outbound radiation has been primarily an endeavor of governments. (Building underground has been an alternative approach.) A future technique at this scale may be to use conductive concrete, originally developed to melt snow. (Preparing the concrete is tricky, so only prefabricated slabs are commercially available at present.)

Wider application of shielding at the level of components and their connecting wires seeks to improve EMC so that each component functions properly. All computers emit RF radiation, and government regulations limit how much radiation is acceptable and where computers may be used. To achieve EMC in components, there are specially designed, conductive enclosures, gaskets, meshes, pipes, tapes, and sprays. The simplest EMC measure is to use shielded cables and keep them separated to prevent crosstalk. Given what was said earlier about nonstop emissions, RF emitters such as mobile phones should be kept away from computers with sensitive data.

Attenuation (lessening) of emissions is measured in decibels (dB). Each 10-dB drop cuts the strength of the signal to one tenth of what it was, so a 20-dB drop means only 1% of the energy is escaping.

A recent discovery, dubbed *Soft Tempest*, provides an inexpensive, partial solution for video display emissions (comparable to attenuation of 10–20 dB). Special fonts, which appear "antialiased" but crisp on the user's screen, are illegible on monitoring equipment because key information about vertical edges is not radiated. GIF (graphic interchange format) versions of such fonts can be downloaded from http://www.cl.cam.ac.uk/~mgk25/st-fonts.zip. See Anderson (2001) for discussions of this and of a perfect software defense against monitoring of keyboard emissions.

Weather Preparedness

Many regions of the world are subject to seasons when monsoons, hurricanes (typhoons), tornadoes, damaging hail, ice storms, or blizzards are more likely to occur, but weather is inherently chaotic. Even if an event arrives in its proper season, that arrival may be unexpected. In general, the larger the scale of the weather event, the farther in advance it can be anticipated. Despite dramatic advances in the accuracy and detail of regional forecasting, the granularity of current weather models does not allow precise forecasting of highly localized phenomena beyond saying, "Small, bad things may happen within this larger area." As the probability of any specific point in that area being hit with severe weather is small, such generalized warnings often go unheeded.

Fortunately, the formation of small, intense weather events can be detected by modern radar, and warnings of potential and imminent danger can be obtained through a variety of means. There are radio receivers that respond specifically to warnings transmitted by meteorological agencies or civil authorities. The Internet itself can be the messenger. One mode of notification is e-mail. Other services run in the background on a client machine, checking with a specific site for the latest information. Some of these services are free (though accompanied by advertising banners). There are also commercial software products and services that give highly detailed predictions in certain situations. For example, one suite of hurricane-related products can predict peak winds, wind direction, and the arrival time of damaging winds at specific locations.

Fitted covers for equipment can be quickly deployed to protect against falling water from a damaged roof, overhead pipe leaks, or sprinkler systems. They can also be used as dust covers when equipment is moved or stored, during construction work, or when the panels of a suspended ceiling need to be lifted.

As noted earlier, lightning can be surprisingly invasive, penetrating where rain and wind do not. Moreover, it does not always hit the most "logical" target, and it can arrive unexpectedly. A bolt was documented to have traveled horizontally 16 km (10 miles) before landfall; it appeared to come out of a blue sky when, in reality, it originated in a cloud hidden behind a hill. In any case, few businesses will be willing to disconnect from the electric grid every time the potential for lightning exists. Consequently, it is essential that a building have a lightning protection system in place and that surge protection be provided for equipment. As a secondary precaution, magnetic media and sensitive equipment should be kept away from metal objects, especially structural steel. On the other hand, storage *within* a metal container affords the same protection that passengers enjoy within the metal body of an automobile; this is called the *skin effect* because the current passes only through the outer skin of the metal. (The rubber tires would need to be a mile thick to provide equivalent protection.)

It is now possible to receive automated alerts regarding impending adverse space weather. The service can be tailored with regard to the means of notification (e-mail, FAX, or pager), the type of event expected (radio burst, geomagnetic impulse, and so forth), and the threshold at which a warning should be reported. See Space Environment Center (2002).

Earthquake Preparedness

Certain regions of the world have a well-known history of frequent earthquakes, and planning for the inevitable is second nature. Complacency prevails where damaging earthquakes strike decades or centuries apart; earthquake survivability features may not be required by building codes (although some cities are waking up to the importance of such measures) or may not be calculated

to be cost-effective. The collapses of the buildings at the World Trade Center had earthquake-like effects on neighboring buildings. (Even the initial crashes registered on seismographs.) Because disasters can occur in anyone's neighborhood, any structure may be subjected to "seismic" forces.

Regardless of construction techniques, how the occupants furnish buildings is largely their own responsibility. Some precautions can be taken with relatively little expense or intrusion to normal operations. Following are three suggestions from Garfinkel (2002) based on the simple principle that objects will move and perhaps fall from high places to lower places:

1. Place computers under sturdy tables, not on high surfaces or near windows.
2. Do not place heavy objects so that they could fall onto computers.
3. Restrain the possible movement of computers with bolts and other equipment.

The first two recommendations also help in case damaging wind (including the force of an external explosion) blows out a window or damages a roof. The last could also serve as a theft deterrent, depending on the type of restraint used. There are also relatively easy ways to secure things other than computers. For example, bookcases can be bolted to walls so they cannot topple, and books can be restrained by removable bars or straps.

Ruggedization of Equipment

With the upsurge in mobile computing comes an increased risk of damage from shock, vibration, dust, water, and extremes of temperature and humidity. One survey found that 18% of corporate laptops in "*non*rugged" applications had suffered substantial damage (averaging about half the purchase price), implying that more people could benefit from tougher equipment. Laptops and other mobile devices can be *ruggedized* by adding characteristics such as the following:

1. having an extra-sturdy metal chassis, possibly encased in rubber;
2. being shock- and vibration-resistant (with a floating LCD panel or gel-mounted hard drive);
3. being rainproof, resistant to high humidity and tolerant of salt fog;
4. being dustproof (with an overlay panel for the LCD screen);
5. being able to withstand temperature extremes and thermal shock; and
6. being able to operate at high altitude.

Touchscreens, port replicators, glare-resistant coatings for the LCD screen, and modular components are available on some models. Some portable ruggedized units resemble a suitcase more than a modern laptop.

Ruggedization techniques can also be used for any computer that must remain in areas where explosions or other harsh conditions may be encountered. Accessories available are ruggedized disk drives, mouse covers, keyboard covers, and sealed keyboards. (Some keyboards can be rolled up.) Some biometric devices can be used in demanding environments.

Redundancy

Redundancy is the safety net for ensuring integrity and availability of resources. Because of the many facets of the computing environment, redundancy takes many forms. The first thing that comes to mind is backing up data. If only a single copy of information exists, it may be difficult, if not impossible, to reconstruct it with complete confidence in its validity. Not to be overlooked are system software and configurations. They should also be backed up in such a way that restarting the system or restoring it to a nominal condition can be accomplished expeditiously.

There are a wide variety of schemes for creating backups. Most are based on some type of high-density tape. Capacities for some are measured in terabytes. The backup procedure can be either manual or automated. The latter approach is safer because it removes the potential for human error in the process, but an automated procedure should issue a notification if it encounters problems while performing its duties. Backups can be made, managed, and used remotely. Some systems allow access to other cartridges while one cartridge is receiving data. Scalability is an important feature available. As mentioned earlier, tapes that are subjected to repeated reuse should periodically be tested and, if necessary, cleaned by a tape certifier.

Backups should be kept at a separate location, preferably far enough away from the site of origin that a single storm, forest fire, earthquake, or dirty bomb could not damage both locations. At a bare minimum, backups should be kept in a fireproof, explosion-resistant safe; it must include insulation so that heat is not conducted to its contents. Backups that are going off-site (perhaps via the Internet) should be encrypted. In all cases, access to backups should be restricted to authorized personnel.

Point-in-time recovery requires not only periodic backups but also continual logging of changes to the data since the last complete backup so that files can be reconstructed to match their last version. Although the need to backup digital information is well recognized, essential printed documents are sometimes overlooked. These can be converted to a more compact medium (e.g., microfilm).

Redundancy in the availability of power can be achieved using a UPS (discussed previously). Some systems themselves have redundant batteries and circuitry. Nonetheless, most UPS systems have backup times designed only to allow controlled shutdown of the system so that no data is lost or equipment damaged. For continued operation during extended blackouts, a backup generator system will also be necessary. It is tempting to place large UPS systems and generators in a basement, but that can backfire if the power outage is concurrent with water entering the building. It is important to anticipate plausible combinations of calamities.

Telephone redundancy has its difficulties. Cellular communications should be available in case wired phone

service to a building is interrupted, but phone systems in general become overloaded and may sustain damage as a result of a major event. Or cellular services could be shut down (as occurred on September 11, 2001, for fear they might be used to trigger bombs). An alternative emergency communication system would be a battery-powered, two-way radio that broadcasts on a frequency monitored by emergency agencies. In any case, RF-emitting devices must not be active near equipment that could suffer from the emissions.

ISP (Internet service provider) redundancy is also complicated. Politically, operationally, and economically, it may make sense to have a single ISP. From the standpoint of robustness, it is better to have at least two service providers and to have their respective cables exit the organization's physical perimeter by different routes (so that any careless excavation cannot damage both lines). Internally, the organization must be able to switch critical services promptly from one provider to the other.

The ultimate redundancy is a hot site, ready to take over operations. This does not need to be owned outright; services of this sort can be contracted.

Sanitization of Media

At some point in time, every piece of storage media of every type will cease to play its current role. It may be reused to store new information, it may be recycled into a new object, or it may be "destroyed" in some sense (probably not as thoroughly as by incineration). If the media is to be used by another individual not authorized to access the old information, the old information must be purged. In the case of recycling or destruction, the original user of the media may assume that no attempt to access the old information will be made after it leaves his or her possession; as was pointed out in the discussion of dumpster diving, this is a foolhardy assumption. Sanitization of media that held sensitive information at any time is the responsibility of its owner.

Printed media holding sensitive information can be shredded. Some shredders are worthless, slicing pages into parallel strips, which can be visually "reassembled." At the other extreme is government equipment that liquefies documents to the point that they cannot be recycled (due to the destruction of the paper fibers). In between are crosscut shredders that produce tiny pieces of documents, a reasonable approach.

For magnetic media, one of the best known vulnerabilities comes from "deleting" a file, which really only changes a pointer to the file. There are commercial, shareware, and freeware tools for (repeatedly) overwriting files so that each byte is replaced with random garbage. Echoes of the original information may remain in other system files, however. Another potential problem is that sectors that have been flagged as bad might not be susceptible to overwriting. Special, drive-specific software should be used to overwrite hard drives because each has its own way of using hidden and reserved sectors.

Even after all sensitive bytes have been overwritten by software, there may still be recoverable data, termed *magnetic remanence*. One reason is that write heads shift position over time, that is, where new bytes are written does not perfectly match where the old bytes were written. Hence the use of a degausser (bulk eraser) is generally recommended. Some models can each accommodate a wide range of magnetic media, including hard drives, reel or cartridge tape, and boxed diskettes. Degaussers are rated in Gauss (measuring the strength of the field they emit), in Oersteds (measuring the strength of the field within the media they can erase), or in dB (measuring on a logarithmic scale the ratio of the remaining signal to the original signal on the media). A degausser generates heat rapidly and cannot be operated continuously for long periods; it should be equipped with an automatic shutoff feature to prevent overheating. Even degaussing may leave information retrievable by an adversary with special equipment. Another suggestion is to grind off the surface of a hard drive. For more information on magnetic remanence, see National Computer Security Center (1991), also known as the Forrest Green Book in the Rainbow Series.

Guidelines for sanitizing write-once or rewritable optical media are not as clear. In theory, even write-once disks can be overwritten, but this is not reliable. Two "folk remedies," breaking the disk or placing it in a microwave oven for two seconds, should *not* be used. Another suggestion, scratching, may be ineffective because there are commercial products and services for repairing scratched disks by polishing. Therefore, if complete destruction of the disk is not possible, it should be ground to the point of obliterating the layer on which the data is actually stored.

For maximum security in recycling or disposing of media, study forensic science as it applies to computing (a separate article), and learn to think forensically—if a government agency could recover information from your media, so could a sufficiently sophisticated adversary.

Physical Security Awareness Training

Because security is everyone's business, education is one of the most important aspects of physical security. It is also cost-effective. Proper practices cannot replace expensive security equipment, but improper practices can negate the value of that equipment. All personnel should be trained how to react in case of a fire, the most likely threat to life in a computing facility. The most important aspect is *practicing* egress procedures. In the areas where total flooding (to be discussed later) is to be employed, occupants of those areas must understand the different alarms, must know how to proceed when the first alarm sounds, and must appreciate the seriousness of that environment. (A short science lesson might help.) All personnel should be acquainted with the location and proper use of portable fire-suppression devices. If more than one type is available, they must know which type is suitable for which kinds of fires. Depending on how many operations are automatic, certain people (enough so that an adequate number are always on duty) must be trained to perform extra duties, including shutting off electricity and natural gas, calling emergency officials, and operating special fire systems (hoses, wheeled portable units, manually controlled sprinklers, etc.).

The variety of possible disasters is so broad (e.g., fallen space debris—with or without radioisotopes), it is impossible to educate employees with regard to every

eventuality. The solution is to teach general principles. In the case of hazardous materials, personnel should just call the proper agencies and get out.

All employees need to know how intruders might enter, how to recognize intruders, and how to react—whom to call and what to do until they arrive. Custodial personnel may need additional training and oversight. They often work at night, a time favored by certain types of intruders. Cleaning crews also are prone to breach security protocols to streamline their work, for example, by leaving offices open and unattended for periods of time. For this reason, education should be reinforced by spot checks to see what is actually going on.

Maintenance and construction workers (whether they are employees or not) must be made of aware of the dangers posed by dust, even from something as simple as accessing the space above a suspended ceiling. When dust-producing activities are anticipated, other employees should know to take precautions, such as installing dust covers on equipment.

All employees who know anything that might be useful to a potential attacker need social engineering awareness training. They should also be educated as to the kind of information that might leak onto a newsgroup bulletin board and why this is bad. For both of these, sample scenarios should be described.

Perhaps the most sensitive area of training regards malicious insiders. Again, sample scenarios can help. Smaller institutions in which everyone knows everyone else are especially likely to have coworkers who are overly trusting of one another. The trick is to preserve the esprit de corps and avoid breeding mistrust among coworkers. The corporate culture should foster "collegial paranoia." Physical security is just another problem that needs to be attacked with teamwork, a highly valued corporate virtue. That means everyone should expect cooperation from everyone else in adhering to physical security protocols. Everyone should believe that an unattended computer is a bad thing. Everyone should expect to be turned down when asking to "borrow" someone else's account; this kind of rejection should not be perceived as a bad thing. (Incidentally, system administrators need to keep in mind that no group of people should be given a common account name and password because this complicates tracing malfeasance to a single person.) Given what has been said about theft of bandwidth and time, appropriate-use policies must be communicated and justified. This is an area where the rules may be less clear-cut than for dealing with colleagues.

Ultimately, the goodwill of employees is invaluable. Managers at all levels must be educated to appreciate the crucial role they play in maintaining an environment which does not turn employees against the organization. Understanding that most attacks are from within is the first step.

REACTIVE MEASURES

Despite the best preventive measures, things will go wrong. Defense in depth requires us to be prepared to react to those calamities. This is most critical when lives are in danger.

Fire Suppression

Fire suppression systems generally release water, dry chemical, or gaseous agents. The release can be from portable devices, from a centralized distribution system of pipes (perhaps with hoses which will be manually directed), or from modular devices in fixed locations. Fire can be extinguished by displacing oxygen, by breaking the chemical reaction, by cooling the fire's fuel below its point of ignition, or by a combination of these.

Any fire in a computing environment should be considered a *Class C* fire because of the presence of electricity. Electrical power should be cut as soon as possible, regardless of whether a conductive fire-suppression agent is used, because any electrical shorting will work against the suppressant. Obviously, automatic fire suppression systems must be able to function independent of the facility's main power supply.

When possible, it is preferable to extinguish a fire immediately with portable extinguishers aimed at the base of the fire before it can grow. Each device should have one or more letters on the label, indicating the class(es) of fires on which it can be used. For most computing facilities, a dry chemical extinguisher rated A-B-C will cover all situations. The dry chemical will leave a residue, but if the fire can be caught early, this is a small price to pay.

Countermeasures must match the potential conflagration, both in quantity and quality. The presence of flammable materials requires greater suppression capacity. In addition, special tools and techniques are needed for special fires. A *Class D* fire (involving combustible metals such as magnesium) requires the application of a metal-specific *dry powder* (so named to distinguish its purpose from that of ordinary dry chemical with B-C or A-B-C ratings). Recently certified, specialized (wet chemical) extinguishing equipment should be installed if there is the potential of a *Class K* fire (involving cooking equipment using oils and fats at high temperature).

Total Flooding with Gaseous Agents

Total flooding seeks to release enough of a gaseous agent to alter the entire atmosphere of a sealed area (with openings totaling no more than 1% of the total surface area of the enclosure). The term *clean agent* is often used to indicate that the gas itself leaves no residue (although its decomposition by-products will). Ordinarily, the air-agent mixture alone would be safe for humans, but fires always produce toxic smoke.

Consequently, the best protocol is to have an alarm continuously announce the impending release of a flooding agent, allow a reasonable time period for personnel to evacuate and seal the area, and sound a second alarm to announce the actual release. Doors must be self-closing and have "panic hardware" for easy exit. Warning signs must proclaim the special nature of the area. Self-contained breathing equipment must be available for rescuing people.

The sudden release of a highly pressurized gaseous agent has several side effects. The gas undergoes a dramatic decrease in its temperature. Reportedly, skin in direct contact with a release could suffer frostbite. Equipment could suffer as well. The force of the exhaust is

considerable and should be taken into account when placing the vents. The noise of a release is loud but not damaging to hearing.

Gaseous fire-suppression systems can be either centralized or decentralized. In the former, a network of pipes delivers the suppressant from a single tank to multiple nozzles operating simultaneously; this is the more traditional and common approach. In the latter, independent units each have a tank, triggering device, and nozzle; they can be equipped for remote triggering or monitoring. Centralized systems are generally custom fitted for a particular installation. Decentralized systems are modular, so there is greater flexibility in placing the individual units or repositioning them (upon expert advice) if the layout of a facility changes. On the negative side, the individual units, being self-contained, are heavier and bulkier than the outlets and pipes of a centralized system. Therefore, they must be supported from a structural ceiling rather than a suspended ceiling. Moreover, each cylinder must be anchored very securely to prevent Newton's Third Law of Motion from turning it into a projectile upon the release of gas. Gaseous agents that have been used in computing facilities include carbon dioxide, argon, nitrogen, halogenated agents (halons), newer replacements for halons, and mixtures of these. (Pure CO_2 at the concentration needed for total flooding is hazardous to humans.)

For decades, the fire-suppression technique of choice in computing facilities was total flooding with Halon 1301 (bromotrifluoromethane or $CBrF_3$). (Halon 1211, a liquid streaming agent, was also used in portable extinguishers.) Because of their ozone-depleting nature, proportionally worse than CFCs (chlorofluorocarbons), halons were banned by the Montréal Protocol of 1987. Disposal and recycling of Halon 1301 must be performed by experts, because it is contained under high pressure. Consult Halon Recycling Corporation (HRC; 2002) for advice and contacts. Although no new halons are being produced, existing systems may remain in place, and the use of recycled Halon 1301 in new systems is still allowed by the protocol (on a case-by-case basis) for "essential" use (not synonymous with "critical" as used by the HRC). Because the world's supply has been decreasing since 1994, a concern when relying on Halon 1301 is its future availability.

Halon 1301's effectiveness is legendary. One factor is its high *thermal capacity* (ability to absorb heat). More important, it also appears to break the chemical chain reaction of combustion. Although the mechanism by which it does this is not perfectly understood (nor, for that matter, is the chemistry of combustion), the dominant theory proposes that the toxins into which it decomposes at about 482°C (900°F) are essential for chemical inhibition.

In low-hazard environments, a concentration of approximately 5% Halon 1301 by volume suffices. Short-term exposure at this level is considered safe but not recommended for humans; dizziness and tingling may result. An even lower concentration is adequate when the Halon 1301 is delivered with a dry chemical that inhibits reignition. Regardless of the concentration applied, immediately after exposure to Halon 1301 (perhaps from an accidental discharge), a victim should not be given adrenaline-like drugs because of possibly increased cardiosensitivity. The real risk comes when fire decomposes Halon 1301 into deadly hydrogen fluoride, hydrogen chloride, and free bromine. Fortunately, these gases, being extremely acrid, are easy to smell at concentrations of just a few parts per million.

In addition to the natural inert gases, there are a numerous replacements for Halon 1301 in the general category of halocarbon agents. Subcategories include: hydrofluorocarbons (HFCs), hydrochlorofluorocarbons (HCFCs), perfluorocarbons (PFCs and FCs), and fluoroiocarbons (FICs). None of these or blends of them seem to be as effective, that is, more of the substance is needed to achieve the same end. The search for better clean agents continues. See National Fire Protection Association (2000) for guidelines regarding clean agents.

Water-Based Suppression

Despite its reputation for doing as much damage as fire, water is coming back in favor. Because water's corrosive action (in the absence of other compounds) is slow, computer equipment that has been sprinkled is not necessarily damaged beyond repair. In fact, cleanup from water can be much simpler and more successful than from other agents. Water also has an outstanding thermal capacity. Misting is now used as an alternative to Halon 1301. The explosive expansion of the steam contributes to displacing oxygen at the place where the water is being converted to steam, namely, the fire. (Steam itself has been used as a suppressant.) Pipes for hose, sprinkler, and mist systems should remain dry until needed to reduce the risk of accidental leakage.

First Response to Other Types of Incidents

One of the most likely incidents demanding an immediate response is an unwanted intruder. In general, it is safer to summon security personnel, particularly if the incident warrants detaining the person for civil authorities. Less likely but potentially more dangerous are incidents involving hazardous materials. It is possible to know in advance precisely which ones are in nearby pipelines and storage facilities, but not which ones pass by on transportation arteries. Therefore, it is essential to know whom to call should a *HAZMAT* (hazardous material) event occur or appear to be imminent. The safest course of action in case of pipeline leaks, derailments, truck accidents, or deliberate attacks is to evacuate immediately unless the substance is known with certainty to be benign.

Because of the tremendous variety of characteristics of modern contaminants, a facility contaminated by chemical, biological, or radiological agents should not be reentered until local authorities and appropriately trained professionals give clearance. Some contaminants, such as sarin gas, dissipate on their own. Some, such as the anthrax spores, require weeks of specialized decontamination. Others, such as radiation, effectively close down an area indefinitely.

Disaster Recovery

Disaster recovery can take as many forms as the disasters themselves. A single event may be handled in different

ways or may require a combination of remedies. Data may be retrieved and equipment rehabilitated on- or off-site. Simultaneously, operations may be (partially) restored on-site or transferred off-site. In most disaster recovery planning (the subject of a separate article), the first priority is maintaining operations or restoring them as soon as possible. There are a variety of services that can be contracted for this purpose. Some are mobile facilities.

We concentrate here on the physical aspects of rehabilitating buildings, equipment, and media. Professional disaster recovery services should always be employed for this purpose. Because such specialized companies are not based in every city, however, their response time does not match that of emergency personnel. Yet for many physical disasters, the first 24 hours are the most important in limiting progressive damage, for example, from water and smoke. Consequently, knowledge of what to do during that crucial time frame is essential. Good references in this regard are McDaniel (2001) and the "What to do in the first 24 hours!" links at the BMS Catastrophe Web site (http://www.bmscat.com/were/press.shtml)

Recovering from Fire Damage

Even when a fire has been extinguished, other problems remain. By-products of the fire, perhaps because of the type of suppressant used, may be toxic to humans or corrosive to equipment. As soon as practical after a fire has been extinguished, thorough ventilation should take place. Only appropriately trained and equipped experts should enter to begin this dangerous procedure. Aside from the initial health hazard, improper procedures may worsen the situation. Active HVAC equipment and elevators might spread contamination to additional areas.

Once air quality has returned to a safe level, resources should be rehabilitated. In some cases, equipment will never again be suitable for regular use; however, it may be brought to a condition from which any important data can be backed up, if necessary. The same is true of removable storage media. Paper documents can be restored provided they have not become brittle.

The combustion by-products most devastating electronic equipment are corrosive chloride and sulfur compounds. These reside in particulate residue, regardless of whether dry chemical (which itself leaves a film) or a clean agent (a somewhat misleading term) was applied. In either case, time is of the essence in preventing the progression of damage. Some types of spray solvents may be used for preliminary cleanup. In the case of fire suppression by water, the procedures outlined below should be followed.

Recovery from Water Damage

The first rule of rehabilitating electrical equipment exposed to water is to disconnect it from its power source. Energizing equipment before it is thoroughly dried may cause shorting, damage, and fire. The second rule is to expedite the drying process to prevent the onset of corrosion. Low ambient humidity speeds drying, whereas high humidity (and, even more so, dampness) speeds the corrosive action of any contaminants. If the HVAC system cannot (or should not) be used to achieve a relative humidity of 40–50%, then wet items should be moved to a location where this can be done. Actively applying heat significantly above room temperature must be done with caution, recalling from Table 1 the temperatures at which damage can occur to media and equipment. Handheld dryers can be used on low settings. An alternative is aerosol sprays that have a drying effect. Even room-temperature air moved by fans or compressed air at no more than 3.4 bar (50 psi) can be helpful. In any case, equipment should be opened up as much as possible for the greatest effect. Conversely, equipment should not be sealed, because this may cause condensation to develop inside. Low-lint cotton-tipped swabs may be used to dab water from hard-to-reach areas.

PHYSICAL ASPECTS OF COMPUTER AND NETWORK SECURITY PLANNING

Computer and network security planning traditionally starts by identifying assets. Physical security planning would best begin before there were any assets to protect. Whereas cyberattacks and cybersecurity have little to do with where resources are located, the earliest stages of physical security planning should consider and dictate location.

Locating a facility in a particular region is usually done with an eye to the bottom line. A variety of regional characteristics influence the difficulty of maintaining physical security and can ultimate affect profit: the availability of electrical power and a skilled workforce; the frequency of earthquakes, hurricanes, tornadoes, or wildfires; and the likelihood of terrorism, civil unrest, or regional conflict. The natural traits will stay fairly constant, whereas the political, social, and economic ones may vary dramatically over time.

Locating a facility at a specific site within a region may have an even more profound influence on total risk. New factors, such as topography and neighbors, enter into the equation at this level. A small difference in elevation can make a big difference where flood plains and storm surges are concerned. Higher terrain may initially look safer than a valley but may be dealt bigger surprises due to steep land gradients. The ground underneath may hold more surprises, such as mine subsidence. Rail lines, major thoroughfares, massive electrical lines, natural gas pipelines, and even major water mains pose potential threats. Adjacent establishments may be high-profile targets, have hazardous operations, or produce abundant electromagnetic pollution. Choosing to have no close neighbors may have long-term consequences if adjoining parcels of land are later occupied by high-risk establishments. Being in an isolated area has implications for emergency services.

Locating departments within a building should ideally influence its design and construction. Critical departments and support equipment (including backup power) should be in the safer areas, not in the basement or on the top floor. Within departments, the most crucial resources should preferably be placed away from windows and overhead plumbing. Safes for any on-site backups should be in windowless, interior rooms with high fire ratings. Flammable and hazardous material must be contained

and isolated to the extent possible. Fire divisions inhibit the spread of fire. Other construction techniques brace for earthquakes or high winds.

Once assets are in place, the physical perimeter of the organization must be defined; beyond some point, the responsibility for physical security switches to others (e.g., ISPs and civil authorities). This footprint (often a collection of widely scattered toeprints), determines where certain physical access controls can be installed.

Physical security doesn't stop at the door. Events outside—riots, dust storms, rolling brownouts—can disturb operations inside. Physical security policies must provide for timely, two-way flow of information (e.g., monitoring of weather forecasts and prompt reporting of internal incidents to relevant authorities).

Moreover, there is a virtual perimeter far more vast and complex than the geographic perimeter. Wherever the organization's employees carry assets, physical security is an issue. Although physical access controls, such as biometric devices on laptops, help, mobile assets are at greater risk and, therefore, in greater need of encryption and redundancy. Crafting and communicating clear, effective policies regarding off-site resources are critical. In the end, the competence and trustworthiness of employees are the best defense.

Even if employees leave all physical objects at work, their knowledge remains with them. The usual nondisclosure agreements must be complemented by policies regarding appropriate usage of newsgroup bulletin boards.

Policies for work-related behavior should address the following:

1. access to facilities and services (when and where who can do what);
2. appropriate use (how each allowed service may and may not be used);
3. integrity of accounts (leaving computers unattended, lending accounts); and
4. data management (backing up files, recycling and disposing of media).

The most ticklish of these is appropriate use. Some employers prohibit even personal e-mail saying, "I have to work late." Others seem not to care about misuse of resources until glaring abuses arise. Neither policy extreme is optimal; research has shown that productivity is actually best when employees are allowed modest time for personal e-mail and Internet access. An alternative to written policy (and some form of enforcement) is to block specific Web sites or to allow only specific sites. The former is inadequate, and the latter is too restrictive in most cases. Yet another alternative is filtering software for Web usage or e-mail. If activity monitoring is used, notification of employees is not legally required. Nonetheless, it is best to spell out both what an employer expects in the way of behavior and what employees might expect with regard to what they may see as their "privacy." In practice, monitoring should be used to control problems before they get out of hand, not to ambush employees. Activity monitoring as described actually covers a small fraction of the spectrum of security-related behavior.

Appropriate-use policy raises issues larger than the impact on profitability. Allowing an organization's resources to be used to illegally duplicate copyrighted material contributes to a large and growing societal problem. There is an ethical (if not legal) obligation to consider not only theft of one's own bandwidth, but also the theft of another's intellectual property.

Every policy needs to be enforced, but the difficulty of doing so ranges from trivial to highly impractical. Whereas compliance in some areas (e.g., periodic changing of passwords) can be enforced automatically, checking to see where passwords have been written down is a completely different matter.

Additional security policies should be written specifically for human resource departments (e.g., background checks for certain categories of personnel), for managers (e.g., activity monitoring protocols), and for IT administrators (e.g., least privilege, to name only one of many).

The final component, as noted before, is education and enlightenment with regard to physical security. Policies cannot work if employees do not understand the policies *and* their rationales. Policies that are considered to be frivolous or unnecessarily restrictive tend to be ignored or circumvented. (Doors will be propped open.) That belief in policies must come from the top. This may require educating and enlightening corporate leaders, who must then lead by communicating down the chain of command their belief in the importance of physical security.

CONCLUSION

Physical security tends to receive less attention than it deserves. Yet cybersecurity depends on it. The two pillars of security must be balanced to defeat malicious insiders and outsiders. Ultimately, physical security is the greater challenge, because nature can be the biggest foe. Physical security involves a broad range of topics outside the normal sphere of IT expertise. Consequently, to obtain the best protection, professionals in other fields should be consulted with regard to fire detection and suppression, power maintenance and conditioning, access to and monitoring of buildings and rooms, forensic science, managerial science, and disaster recovery. A basic understanding of how these areas relate to physical security facilitates communication with consultants. Combining information with the imagination to expect the unexpected leads to better physical security planning and practice.

The scope of physical security is wider than is immediately evident. It concerns an organization's resources, wherever they go. An asset often forgotten is employees' knowledge. Equally important are their intentions. Thus, physical security involves everyone, all the time. It relates to intangibles such as trust and privacy, and it must look inward as well as outward.

GLOSSARY

Class A fire Fire involving ordinary *combustibles* (e.g., wood, paper, and some plastics).

Class B fire Fire involving *flammable* or *combustible* liquid or gas (e.g., most solvents).

Class C fire *Class A* or *B* fire amid energized electrical wiring or equipment, which precludes the use of extinguishing agents of a conductive nature (e.g., water or foam).

Clean agent Gaseous fire suppressant that technically leaves no residue; residues will result when the agent breaks down under the heat of combustion.

Combustible Capable of burning at normal ambient temperature (perhaps without a flame).

Degausser or **bulk eraser** Alternating current-powered device for removing magnetism (*Degausser* is often applied specifically to wands that rid cathode ray tube monitors of problems displaying colors. The latter term indicates that data is wiped en masse rather than sequentially.)

Electrical noise electromagnetic interference, especially interference conducted through the power input, or minor *spikes*.

Electromagnetic interference (EMI) Undesired electrical anomalies (imperfections in the desired waveform) due to externally originating electromagnetic energy, either conducted or radiated.

Flammable Capable of burning with a flame; for liquids, having a flash point below 38°C (100°F).

Halon or **halogenated agent** *Clean agent* formed when one or more atoms of the halogen series (including bromine and fluorine) replace hydrogen atoms in a hydrocarbon (e.g., methane).

Heating ventilation air conditioning (HVAC) Equipment for maintaining environmental air characteristics suitable for humans and equipment.

Line filter Device for "conditioning" a primary power source (i.e., removing electrical noise).

Radio frequency interference (RFI) Sometimes used as a synonym for *EMI,* but technically the subset of EMI due to energy in the "radio" range (which includes frequencies also classified as microwave energy).

Sag or **brownout** Drop in voltage.

Smoke Gaseous, particulate, and aerosol by-products of (imperfect) combustion.

Spike or **transient** or **transient voltage surge (TVS)** Momentary (less than 1 cycle) increase in voltage.

Surge Sudden increase in electrical current; also used for *spike,* because the two often arrive together.

Tempest or **compromising emissions** Electromagnetic emanations from electrical equipment that carry recoverable information, popularly referred to by the code word for a U.S. government program to combat the problem.

Uninterruptible power supply (UPS) Device to provide battery power as a backup in case the primary source of power failures.

CROSS REFERENCES

See *Computer Security Incident Response Teams (CSIRTs); Disaster Recovery Planning; Guidelines for a Comprehensive Security System.*

REFERENCES

Anderson, R. (2001). *Security engineering: A guide to building dependable distributed systems.* New York: Wiley.

Chomerics. (2000). *EMI shielding engineering handbook.* Retrieved June 19, 2002, from http://www.emigaskets.com/products/documents/catalog.pdf

Cote, A. E. (Ed.). (1997). *Fire protection handbook* (18th ed.). Quincy, MA: National Fire Protection Association.

Garfinkel, S., with Spafford, G. (2002). *Web security, privacy, and commerce.* Sebastapol, CA: O'Reilley & Associates.

Halon Recycling Corporation (2002). *Halon Recycling Corporation homepage.* Retrieved June 19, 2002, from http://www.halon.org

Hartke, J. (2001) *Measures of CD-R longevity.* Retrieved March 3, 2003, from http://www.mscience.com/longev.html

International Advisory Committee for the UNESCO Memory of the World Programme staff (2000). *Memory of the world: Safeguarding the documentary heritage.* Retrieved June 19, 2002, from http://webworld.unesco.org/safeguarding/en

International Commission on Non-Ionizing Radiation Protection (1998). Guidelines for limiting exposure to time-varying electric, magnetic, and electromagnetic fields (up to 300 GHz). *Health Physics,* 75(4), 494–522. Retrieved March 3, 2003, from http://www.icnirp.de/documents/emfgdl.pdf

International Electrotechnical Commission (2001). *Information technology equipment-safety—part 1: General requirements [IEC 60950–1–Ed. 1].* Geneva: International Electrotechnical Commission.

McDaniel, L. D. D. (Ed.). (2001). *Disaster restoration guide for disaster recovery planners (revision no. 10).* Fort Worth, TX: Blackman-Mooring Steamatic Catastrophe.

McNamara, J. (2002). *The unofficial tempest information Page,* Retrieved June 19, 2002, from http://www.eskimo.com/~joelm/tempest.html

National Computer Security Center (1991). *A guide to understanding data remanence in automated information systems, version 2* [NCSC-TG-025]. Retrieved June 19, 2002, from http://www.radium.ncsc.mil/tpep/library/NCSC-TG-025.2.pdf

National Fire Protection Association (1999). *Standard for the protection of electronic computer/data processing equipment* (NFPA 75, 1999 ed.). Quincy, MA: National Fire Protection Association.

National Fire Protection Association (2000). *Standard for clean agent fire extinguishing systems* (NFPA 2001; 2000 ed.). Quincy, MA: National Fire Protection Association.

Skoudis, E. (2002). *Counter hack: A step-by-step guide to computer attacks and effective defenses.* Upper Saddle River, NJ: Prentice Hall PTR.

Space Environment Center (2002) *Space Environment Center space weather alerts.* Retrieved March 3, 2003, from http://www.sec.noaa.gov/alerts/register.html

Politics

Paul Gronke, *Reed College*

INTRODUCTION

Holding torches to light the night sky in October 1876, nearly 4,000 people rallied around a temporary platform in New Haven, Connecticut's sixth electoral ward. One hundred twenty-two years later, nearly 2 million "hits" were recorded on the "Jeb Bush for Governor" Web page, 4,000 Wisconsin citizens signed up for e-mail "listserv" distribution of information about Russell Feingold's (D-WI) Senatorial campaign, and more than 14,000 users posted messages on an electronic bulletin board maintained by the campaign of Jesse "The Body" Ventura (ex-wrestler, talk show host, and current governor of Minnesota). The 1998 election was heralded as the first to demonstrate the potential of the "e-campaign."

By the 2000 campaign, presidential candidate John McCain raised $500,000 in a single day over the World Wide Web. National voter information portals reported hundreds of thousands of hits daily as the election approached, and on election day, governmental sites with real-time election results experienced daily hit rates of 75,000 (Dallas) to 1,000,000 (Washington Secretary of State) on election day (Sarkar, 2000). And when the 2000 presidential contest was thrown into doubt, nearly 120,000 users *per hour* bottlenecked the Florida Secretary of State's Web site. Clearly, e-politics is here to stay.

However, just like the old rules of the stock market, many of the old rules of politics have proved to be surprisingly resilient. Even before the January 2001 presidential inauguration, many of the major politics "portals" had shuttered their electronic doorways or were undergoing strategic makeovers. Media companies that had spent millions of dollars developing an online presence were finding that Internet news sites not just failed to make money but were major sources of revenue loss (Podesta, 2002). Internet connectivity rates had flattened. Clearly, e-politics is off in the distant future.

The reality lies somewhere between these two extremes. The rapid penetration of electronic mail and World Wide Web access into homes and offices, the proliferation of Web sites, and the emergence of the Internet as a new forum for communication present vast new opportunities for citizen participation in the political process. Traditional—and increasingly nontraditional—political organizations (candidate campaigns, political parties, and interest and activist groups) cannot ignore the power of the Internet to *mobilize* citizens.

This chapter will review the impact of the Internet on political participation, using the rational choice model of participation as a lens. According to the rational choice theory of participation, unless individual citizens, after assessing the costs and benefits of political action, find it in their self-interest to participate, they will decline to do so. Although the Internet may lower one cost of participation—easy access to information—the glut of information on the Internet may increase the costs of selection and comprehension. The result may be that citizens will be overwhelmed, continuing to feel that politics is distant, complicated, and marginal. Thus, many citizens continue to have little motivation to get informed and participate. There is little indication that e-politics will change this in the foreseeable future. This same "rational choice" perspective, however, points to those actors and organizations that do benefit directly from politics: political candidates and parties, interest and lobbying groups, and activist organizations. The Internet has had, and will continue to have, its greatest impact as a tool for mobilization efforts by political organizations. In the following sections, I provide a more detailed summary of the rational choice model of political participation, followed by an analysis of how the Internet may change the logic of participation for individuals, and close by extending the review to cover political organizations, parties, and the mass media.

"MACHINE" POLITICS IN AN ELECTRONIC AGE: WHO IS BEING SERVED?

The old political machine, symbolized by Tammany Hall and Boss Tweed of New York or Richard Daley of Chicago,

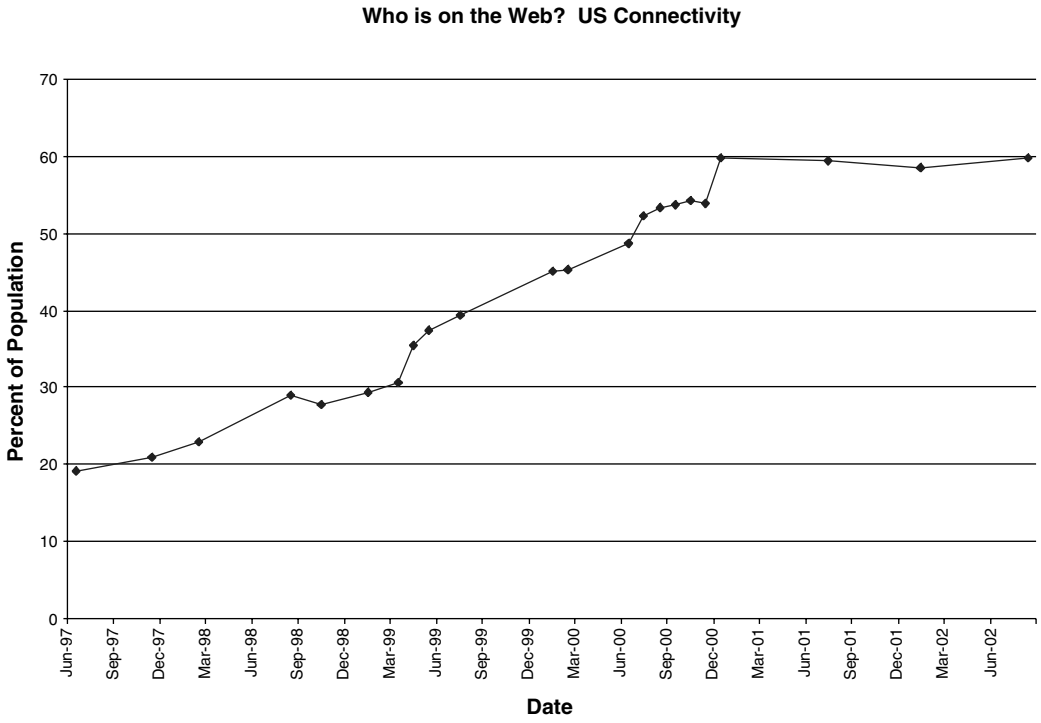


Figure 1: Who is on the Web in the United States? (Data source: NUA Internet Surveys.)

lowered transaction costs for new immigrants and poorly educated urbanites, provided jobs and social welfare (via the patronage system), and encouraged political involvement. This is why, in some quarters, "boss politics," although corrupt by modern standards, is celebrated as a reasonable adjustment of the political system to an undereducated, rapidly urbanizing population.

Is it accurate today to refer to a new "political machine"? Today's political machine is the personal computer, powered by the Internet. Many trumpet the political potential of the Web-connected PC for many of the same reasons that some celebrate the old political machine. The PC and the Internet, they argue, will lower the costs of political information and involvement, make politics more relevant to our daily lives, and consequently substantially increase rates of political participation. The rapid growth of the Internet means that it is far too important for any political candidate or organization to ignore. As shown in Figure 1, Internet penetration rates in the U.S. have climbed dramatically over the past decade and are currently estimated at 60% (though showing little growth in the past year). Perhaps most importantly, the more wired segments of the population—those with higher levels of education, income, and occupational status—are the same segments who are more likely to volunteer, donate money, and vote (Bimber, 2002; Davis, 1999; Rosenstone & Hansen, 1993). A significant proportion (35%) of Americans report going on the Internet at least once a week to get news, although, parallel to penetration rates, this proportion has slowed significantly from its rapid growth in the late 1990s and still lags far behind traditional media sources (Pew Center, 2002). Users with high-speed connections—currently estimated at 21% of U.S. users—report far higher rates of Internet utilization for newsgathering (Horrigan & Rainie, 2002). The Internet is clearly a mass medium for communication.

International Internet penetration rates, however, although they continue to climb rapidly, remain below 10% (NUA Internet Surveys, 2002). As Pippa Norris has shown, this difference means that, except for a few more highly connected European countries, e-politics will remain a distinctly American phenomenon (Norris, 2001).

The new political machine holds the potential for a more egalitarian, democratized, and decentralized political system, whereas the old political machine was the very essence of centralized political control. The machine metaphor is appropriate, furthermore, because it focuses our lens on the area where the Internet has already had, and is likely to continue to have, its greatest impact—on the ability of political elites and organizations to communicate with, mobilize, and potentially control public attitudes and political activities. The Internet has become a central tool for mobilization efforts by political organizations. The rapid penetration of electronic mail and World Wide Web access into homes and offices, the proliferation of Web sites, and the emergence of the Internet as a new forum for communication present vast new opportunities for citizen participation in the political process. The Internet's potential to broaden and increase participation by changing the behavior of individual citizens, however, runs squarely into one of the most widely recognized social dilemmas: the logic of collective action.

RATIONAL CHOICE AND DEMOCRATIC PARTICIPATION

In *Strong Democracy*, political philosopher Benjamin Barber argues that neighborhood assemblies and town

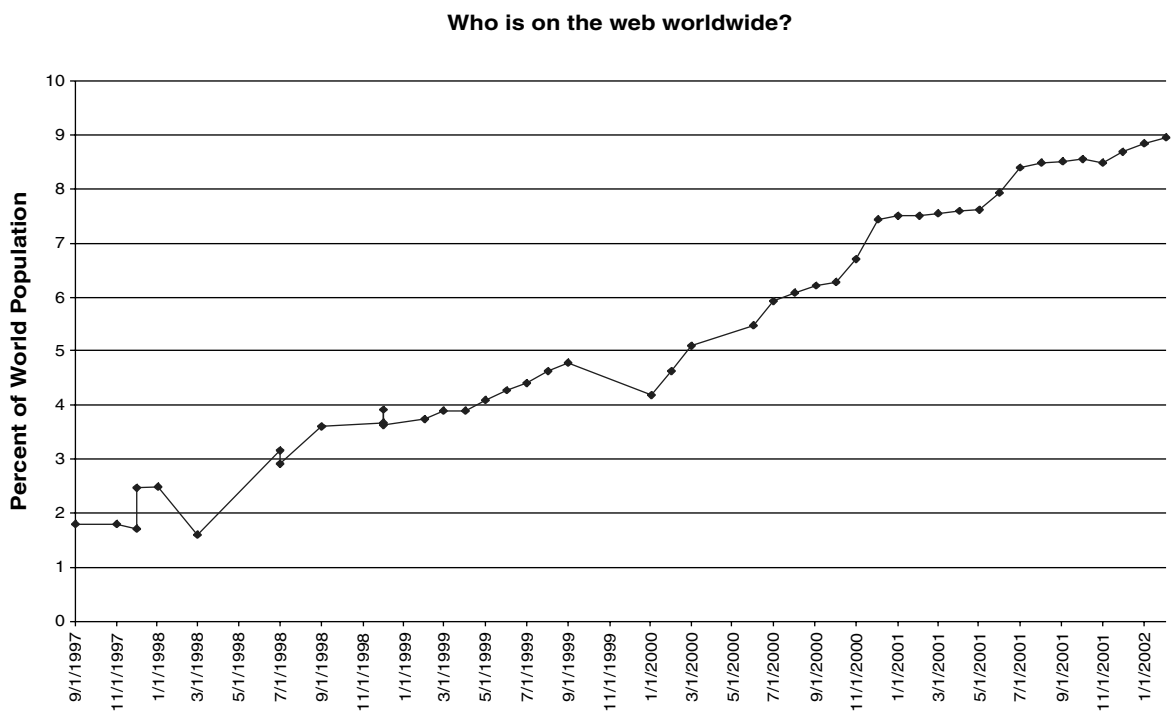


Figure 2: Who is on the Web worldwide? (Data source: NUA Internet Surveys.)

meetings are necessary to create a democracy that relies upon what he calls "strong talk," a democratic community relying upon increased political participation via public discussion and debate (Barber, 1984). Barber addresses the problem of the "zookeeper" mentality of liberal democracies: a system that acts more to *defend* individual preferences and liberty from one another than *promote* shared commitments and civic engagement. The critical missing element, Barber believes, is greater participation. Citizens in most liberal democracies are only "free" every 2, 4, or 6 years—only when they vote.

Whether or not we agree with Barber, few would assert that greater civic participation poses a problem for republican democracy. Though James Madison argues in *Federalist 10* (Hamilton, Madison, & Jay, 1961) that the public opinion of a majority must be filtered by a republican government, nearly everyone agrees that greater involvement in the political and civic sphere adds to the credibility of liberal democracy and that current levels of disengagement in the U.S. are a serious area of concern (Putnam, 2000). However, Barber's strong talk, Putnam's "social capital," and other participation-inducing devices have always encountered problems with real world application: the seeming irrationality of political participation.

Within political science, the dominant perspective for understanding political participation is *rational choice*. According to this view, a rational individual chooses whether to engage in political activity (writing a letter, joining a protest march, voting, etc.) only if the benefits exceed the costs. The argument is deceptively simple, but leads to powerful conclusions:

$$\text{Participate (e.g., Vote) only if } Probability * Benefits - Costs > 0.$$

Verbally, this equation tells us that individuals engage in a particular political act if the *benefits* (say, a particular candidate winning office) exceed the *costs* of participation. Even stated this way, ignoring the other elements, participation looks irrational. The direct benefits to most individuals of, say, George Bush winning the presidency are quite low. These are quickly overwhelmed by the costs of being informed, registering to vote, and actually getting to the polling place and casting a ballot.

The problem becomes insurmountable when we add the *"probability"* term. This term captures what social scientists refer to as the "collective action problem." An election outcome, such as a Bush victory, is a "public good." Public goods, such as clean water or clean air, are defined as goods that everyone can enjoy, regardless of whether or not he or she helped provide the good. An election outcome is a "good" (or "bad" for those on the losing side) which we "enjoy" whether or not we voted. Thus, unless we believe that our single vote will be decisive in the outcome—represented as *"probability"* above—then we are better off staying at home. In most elections the value of *probability* is vanishingly small. The rational citizen should not vote, and certainly should not engage in Barber's strong democracy. This is, of course, the Achilles heel for this theory, because many people *do* vote. As a consequence, some scholars have posited a "consumptive" benefit to participation (a *Duty* term), something that we enjoy whether or not our candidate wins. Although for some, the inclusion of *Duty* solves the puzzle of participation, for others, this reveals the poverty of this approach to political action. For a summary of the rational choice theory of voting, see Aldrich (1993). For a critique of this viewpoint, see Green and Shapiro (1996).

Regardless of the debate, the fact remains that the "equation of political participation" provides a structured

way to think about the impact of the Internet on politics and political action. In general, early commentaries assumed that the Internet would work its wonders on the cost side of the equation, making it easy and cheap for citizens to learn about candidates, and allowing citizens to personalize their Internet experience, so that a participatory revolution would result. These early analyses failed to take into account the fundamental barrier to participation: interest and motivation. We are already buried under an avalanche of political information; increasing the flow will only make it harder to manage the "information tide" (Graber, 1984, 2001). There is little indication, at present, that the Internet has significantly lowered the costs of participation (Davis, 1999).

But the Internet may work changes in the future. The Internet might inflate perceived benefits, if it provided a way for candidates and parties to contact voters and let them know about the advantages of one party over another. The Internet could allow citizens to see interests where they did not exist before, by allowing the creation of "virtual communities" of interest (Davis, 1999, Chap. 6; Turkle, 1997; but see also Bimber, 1998, for a cautionary view). Or it may provide an avenue for organizations to encourage political participation as an act of civic duty. This may mean that mobilization efforts will be cheaper and easier. Finally, it is possible that, by disseminating more accurate information on the relative support for each candidate, the Internet could lead to more precise estimates of "*probability,*" most likely depressing levels of participation. I examine each of these possibilities below.

A second theory, the *institutional model* of politics, dovetails nicely with this model of participation. Political action does not occur in a vacuum: individuals are embedded within a larger set of social and political institutions. Intermediary organizations, such as interest groups, political parties, and the mass media, communicate the preferences of the mass public to governmental actors, educate the mass public about the activities of government, and mobilize the public to participate in politics (Verba, Schlozman, & Brady, 1995; Rosenstone & Hansen, 1993). In an institutional model of politics, special interests, lobbying groups, "issue publics," and political elites are important engines of political change, with the mass public primarily choosing among the contestants at election time. With respect to the Internet, the institutionalist model turns us away from the mass public, and instead asks how the new tools of e-politics may have strengthened or weakened the influence of pre-existing intermediary organizations and possibly allowed new organizations to enter the fray.

Second, the institutionalist model highlights the importance of *political information* for understanding political power and influence. Whether elites control the mass public or vice versa, the primary point to remember is that the cost, accessibility, and accuracy of political information are a key part of democracy, just as obviously, information flow is the *sine qua non* of the Internet. Beyond its role as a tool for intermediary organizations to mobilize the public and influence the government, the Internet could provide a way for citizens to influence government *directly*, bypassing intermediary institutions.

To summarize, the political world consists of the mass public, elites, and pre-existing political institutions. A careful survey of politics must consider the motivations and interests of each set of actors in the political process if we want to understand how a new and potentially revolutionary medium such as the Internet may change the political world. Although the Internet may not change politics in one realm (e.g., it is unlikely to fundamentally change citizen interest or perceived benefits from participation in politics), it could provide invaluable tools in another realm (e.g., making it far easier to raise money, recruit volunteers, and mobilize voters).

THE MASS PUBLIC

Lowering the Costs of Participation via Low-Cost Computing

In the poorest sections of New York City and in the Indian reservations of Arizona, most households deem themselves lucky to have a telephone, much less a computer with access to the Internet. For all of its promise as a mobilizing force, the World Wide Web is simply useless in such places today. Before a move occurs to widespread online voting or Internet domination of political discussions, a larger portion of the population must have access to personal computers than is the case today. Luckily, the price of a PC has declined in a relatively predictable manner for almost two decades in concert with a steady rise in computing capabilities. Such trends will almost undoubtedly continue for at least several years into the future.

Several characteristics of personal computers improve so steadily as to have "laws" coined based on their progress. For example, "Moore's Law" states that the number of transistors on a microchip doubles each year (Moore, 1965). Likewise, the cost per megabyte of DRAM falls by an average of 40% each year (Hennessy & Patterson, 1990, p. 8). Because a recent estimate of low-end machines placed the percentage of material costs attributable solely to DRAM at 36% (the highest ratio of any component), there is significant room for improvement despite the seeming bargains found in retail stores. Gains made in video systems and monitors (summing to another 36%) will also contribute strongly. As long as the price for which a machine is sold does not fall below its material costs and construction overhead, computer manufacturers will attempt to sell PCs in bulk and profit from volume.

The commoditization of the Internet PC may someday make the machine as widespread as the telephone or the television. When the computer achieves such household status, it indeed seems likely that it will become the primary means by which political information is gathered if not the primary method by which political participation takes place. But if previous telecommunications revolutions have not transformed political participation, why will the Internet?

New Tools for Political Learning and Interaction

Information is the *sine qua non* of the Internet. In the near future, changes in technology may lower the costs of

participation in forums such as Barber's electronic town hall, especially as a faster and more interactive Internet allows more flexibility and greater ease of use. Two areas of enhancement in particular, the increasing use of audiovisual components in Web pages and the increasing spread of residential high-speed Internet connections (via both cable and phone lines), should allow citizens to participate in virtual local government assemblies or neighborhood forums with the same effectiveness and clarity as if all participants had actually gathered in the same physical meeting space. Thus, participation itself might have a tangible benefit—entertainment and enjoyment—even if it does not translate into direct "benefits" from a political outcome. In the next section, we chart the advance of these technologies and speculate as to what effects these advances might have on political participation, town hall meetings, interactive government, and online deliberation and discussion.

AudioVisual Services

Like the transition from newspaper to radio and then to television during the first half of the 20th century, the Internet has undergone in the past five years a transition from a primarily text- to image-based form of communication. Increasing bandwidth, short attention spans, and a need to differentiate a site from its competitors have driven this increase in audio and video online. As was the case with the first Web pages, audiovisual plug-ins began to appear on large commercial sites with plentiful resources, as well as on the Web sites of educational institutions. And just as the second generation of HTML editors made writing a Web page as easy as typing in a word processor, the newest generation of editors is slowly democratizing these technologies by lowering the learning curve required to incorporate them. The first decade of the 21st century will likely see the reinvention of the Web as a multimedia communications center.

The move to augment text with voice has been slow (in Internet time) but steady. Common e-mail applications such as Eudora and Outlook have for several years included audio plug-ins, allowing users of a single application to exchange messages in this way. The lack of an industry standard has slowed the popularization of voice messaging, allowing it to be overshadowed by more recent innovations such as Web telephony, music downloads, and even online wake-up calls. Although many netizens are just becoming accustomed to exchanging electronic voice mail and publishing musical compositions online, power users have begun to tinker with online video. The ability to publish home videos and self-produced animations, combined with the growing popularity of DVD recorders and other such devices, opens up doors previously unimaginable.

As these multimedia tools are simpler to use, and broadband connections become more common, multimedia creations will become commonplace. This is already evident at political Web sites: a study by Kamarck and Nye (1999) found that, even by 1998, most Congressional candidates' Web sites incorporated audiovisual, multimedia, and interactive services as part of their content (see also Wu, 1999). The move to a more visually compelling Internet presages the day when Web-based political communications will rival those currently available only on television and radio.

High Speed Internet for Everyone?

A precursor to the use of the Internet as a visually compelling medium for political information gathering, however, is a broadband connection. Although multimedia-enhanced newsgroups, streaming discussion groups, and even searchable archives of campaign videos are already available, experiencing them becomes an almost painful experience without sufficient bandwidth. On the client side, the race between cable modems and ADSL connections has brought the price of both services within reach of those of modest incomes, although not as inexpensive as was first hoped by Congressional advocates of telecommunications reform in 1996 (as illustrated in the debate over the 2002 Tauzin–Dingell Broadband Deployment Act).

Whether via coaxial cable or twisted-pair copper, nearly 25 million Americans have already found their way onto the high-speed Internet (Horrigan & Rainie, 2002). As the technologies mature, monthly fees should continue to fall and the move to ADSL and cable will accelerate.

Will broadband make a difference in the political impact of the Internet? Early indications are that broadband access will be decisive. Horrigan and Rainie's recent study, undertaken as part of the Pew "Internet and American Life" project, indicates that broadband "transforms" the Internet experience. Broadband users are far more likely to access the Internet on a daily basis and are two to three times as likely to use the Internet to collect news, product, travel, and educational information. Most importantly, for anyone who subscribes to Barber's model of a "strong" democracy consisting of active, participatory, and community-minded citizens, broadband users are far more likely to be *content providers*, setting up Web pages, storing photos online, and sharing information with others (Horrigan & Rainie, 2002, pp. 12–14). Again, for these users, the direct benefits of "participating" (in this case, setting up a Web site) seem to exceed the costs. However, this same study shows that broadband access is heavily skewed toward the same groups that have been traditionally advantaged in the political realm—well-educated, higher income, and now technologically savvy segments of the population. Far from democratizing, the Internet might even exacerbate income, educational, and racial disparities.

A Case Study in the Internet as a Tool of Mass Participation: E-voting

In the March 2000 Arizona Democratic Presidential primary, the first-ever binding Internet vote in a Presidential primary, a vast number of Arizona Democrats participated relative to previous elections (Chiu, 2000). Many speculated that Internet voting mobilized the electorate and provided lower costs to voting—thus creating a higher turnout. If we believe that some of the high turnout for Arizona's primary can be attributed to Internet voting, than electronic referenda could gain support as an untapped resource for furthering political participation.

Online voting could have a substantial impact on the greatest flaw of the suffrage: decreased turnout. In

addition, online voting might lower the cost of voting for those without adequate transportation. Though this would involve a significant change in the Internet usage rate among the poor in the United States, this mobilizing effect remains a possibility. If universal Internet access became a reality, the increased percentages of racial minority voters could help assuage the concerns of critics concerned about the protection of racial minority interests in an election. Finally, if electronic balloting is preceded by widespread "strong talk" and/or "deliberative polls" (Fishkin, 1991), this ongoing democratic conversation could substantially improve the quality of democratic dialogue and decision-making.

On the other hand, critics of Arizona's election, such as the nonprofit Voting Integrity Project and the National Commission on Federal Election Reform, believe online voting is not currently technically feasible (or, if feasible, would require violations of privacy that would be objected to by most Americans) (Phillips, 1999). Internet voting could also lead to discrimination against those without access to the Internet and opens up the possibility of election fraud (Phillips, 2000). Others argue that it erodes civil society by individualizing what used to be a community-based participatory act (Hansen, 2001). Though it seems evident that low-cost computers and Internet access might someday soon be universally available, that day is not yet here. Activist organizations and scholars continue to criticize online voting for its promotion of unequal opportunities in a participatory democracy.

In sum, low-cost computers and universal Internet access have the *potential* to revive the movement toward national referenda (Barber, 1984, p. 281), enhance democratic discussion, and increase voting turnout. However, the *reality* is far less certain. Most importantly, the Internet could make it even less likely that an individual will find it rational to participate. Anything that increases the size of the electorate will simultaneously decrease the probability that an individual vote will be decisive. More likely, however, the Internet will provide new channels for political organizations to *mobilize* citizens and increase participation. Only time will tell whether enhanced mobilization will equalize political influence, or only exacerbate existing inequalities, as current mobilization efforts do (Rosenstone & Hansen, 1993). Internet voting is coming to a computer near you in the next decade, but likely later rather than sooner.

The Mass Public in a Wired World: Old Wine in New Bottles?

In 1984, Barber imagined televised town meetings, which could allow citizens to become more involved in civic affairs. Today, his vision of a televised town hall could evolve into a teleconferencing meeting that could allow thousands to participate. The key is the technological capability and bandwidth to simultaneously stream unlimited numbers of audio and visual inputs into one electronic meeting room.

This is an alluring vision, but what is the reality of participation via the Internet? In electronic town halls, each participant in the electronic town hall must have the technological capacity—and desire—to participate. If town halls were *exclusively* electronic, than the universal availability of high-speed Internet service and fast computers would be a necessity in order to avoid barriers to participatory democracy. And if just the most politically interested entered this conversation, the dialogue would be just as biased toward certain segments of society as it was in the pre-Internet period.

Unfortunately for optimists predicting a participatory revolution fueled by the lower communication costs of the Internet, few studies indicate that the Internet will have any mobilizing effect—a force that makes political activists out of current nonactivists (e.g. Davis, 1999). Although the Internet may have reduced some costs of getting informed, it has not, as yet, increased citizen interest and motivation.

Furthermore, it is not clear that the Internet will necessarily serve as a force for citizen control. Lawrence Lessig notes that although the Internet as currently constructed is a venue for democratized information flow, there is no reason that it needs to be constructed in this way. It is just as likely, via control of *code*, that elites, corporations, and governments will use the Internet to monitor and control our daily lives. Our cyberidentities and cybercommunication are ultimately subject to the restrictions placed upon us by those who write the software and manufacture the hardware. In Lessig's view, the Internet may just as likely strengthen the hands of large, centralized corporations and governments. Witness the Communications Decency Act (CDA) in the U.S. and the many efforts by other governments (e.g., China, Singapore) to control the flow of information available on the Web (Lessig, 1999, 2002). Cass Sunstein points out that the very element of the Internet that many celebrate—individualized control over the interactive experience—could hamper political and civic life. A healthy democratic polity requires that we confront viewpoints that are opposed to our own. A personalized Internet experience, however, could result in reading only news that we agree with, participating in discussion forums with like-minded partisans, and learning about candidates for whom we are already inclined to vote (Sunstein, 2002). And there is no guarantee that the interest groups, news organizations, and other well-funded organizations that sponsor such "forums" or town meetings will allow dissenting voices. What sort of democratic polity would result from such a "personalized" world of political interactions? According to Bruce Bimber, the most likely outcome is "accelerated pluralism," where America's already fragmented political community becomes even more divided (1998). This is a worrisome vision.

Finally, would an electronic town hall be more effective and mobilize new participants in the political arena? A February 2000 article in *The San Francisco Chronicle* detailed the efforts of ActionForum (www.actionforum.com,) a new Web site in Berkeley, California, designed to promote increased civic and political discussions. The city wanted to boost its civic participation because only 418 of its 108,000 citizens spoke at city council meetings in 1999. The site, which consists of an upscale newsgroup bulletin board, received 75 postings in its first month of use. The *Chronicle* reported that most of the authors were familiar faces on the political participation circuit (Holtz, 2000).

Those who had the civic sense or personal motivation to participate via traditional methods simply reappeared in the new forum.

The general mobilizing effects of teleconferencing or high-speed Internet access seem nearly impossible to prove. Most studies to date, including the Pew Research Center for the People and Press's 1996 and 2002 studies, conclude that the Internet, thus far, acts more often as a "re-enforcement" agent, which merely changes the venue in which political participation takes place (Pew Research Center, 1996, 2002). Richard Davis points out that most political activities on the Internet are electronic analogs of activities carried out via older media such as television, newspaper, radio, and mail. In fact, Davis further argues that the Internet could lead to greater political apathy by providing a politically apathetic generation of young Americans with individually tailored, nonpolitical news (see Sunstein for a contrary viewpoint). However, the specific mobilizing or re-enforcement tendencies of high-speed Internet connections and audiovisual enhancements cannot yet be determined, because no strong evidence for either argument yet exists. These conclusions echo the findings of scientific studies of participation conducted over the past 30 years. Participation is skewed towards the well off, well educated, and politically motivated (Rosenstone & Hansen, 1993; Verba et al., 1995). New modes of participation, such as the Internet, are unlikely to change this state of affairs.

POLITICAL INSTITUTIONS: THE INTERNET AS A TOOL OF MOBILIZATION

"Intermediary" organizations—such as political parties, candidate organizations, interest groups, and the mass media—are not hampered by the logic of collective action or by the irrationality of political action. Quite the opposite: for these organizations (as well as for political candidates and entrepreneurs), the benefits of political activity outweigh the costs; otherwise they would not exist (Olson, 1971; Rosenstone & Hansen, 1993). It is no surprise, then, that it is among these pre-existing organizations that the Internet has proved to be a truly revolutionary force. The Internet is a tool to more efficiently and more cheaply communicate their positions to the mass public and mobilize citizens for political action. In this respect, then, the Internet will change mass democracies, not by transforming the public, but by transforming elites, making it easier, cheaper, and quicker for candidates to mobilize supporters and for interest groups to recruit members. Note that "elites" refers to a far broader segment of the population than just the moneyed or politically powerful. It may also include antiestablishment groups, such as the WTO activists, who very successfully organized via the Internet.

Campaign Use of the Internet

In the years between 1992 and 1996, campaign Web sites went from novelty to necessity. In 1996, Bob Dole concluded the second of his Presidential debates with Bill Clinton by plugging his campaign Web site. The era of the campaign Web site as an integral part of the campaign process had begun. Easy-to-follow guidebooks for setting up a campaign Web site are readily available (Democracyonline.org's "Online Campaigning: A Primer") and the Federal Election Commission has clarified the place of campaign Web sites in the campaign finance system (Corrado, 2000). By the 2000 campaign, virtually every candidate for federal office and many state and local candidates had a campaign Web site.

Recent elections have shown that the Internet is a new and important source of campaign funding (Thornburg, 2001). In the 2000 presidential election, Republicans George W. Bush and John McCain and Democrats Bill Bradley and Al Gore used the Internet to solicit funds, with McCain raising more than $500,000 the first day his Web site came online. Internet donations are a small part of overall campaign funding but they have the potential to become much larger, because of low cost and ability to target supporters. In most forms of solicitation, the more people the candidate wishes to reach, the higher the cost. However, there is very little difference in cost for a candidate having 10 or 100,000 people view a Web site. Similarly, the Internet provides a way for candidates to better target supporters. An example would be e-mail lists; they can be set up to better find those who support the candidate and are likely to give him or her money. Because it is so cheap and so effective, the Internet will make it easier for less well-known candidates, parties, and groups to make their voice heard in elections.

Early campaign Web sites, in many cases, were nothing more than electronic brochures, Web-formatted versions of the same leaflets that campaign volunteers had previously passed out on street corners. In short, politicians failed to produce "sticky" Web sites that increased the amount of time spent at a site and the frequency with which users returned to that site. In their study of both political and e-commerce Web pages, James Sadow and Karen James (1999) found that the political sites in 1996 and 1998 lacked the interactive elements that would make the sites more effective in drawing surfers and retaining interest. Citing several studies of e-commerce sites, their study claims that greater interactivity, defined as "the extent to which users can participate in modifying the form and content of a computer mediated environment in real-time," leads to more positive attitudes about specific Web sites and a greater ability to attract consumers (also see Ariely, 1998; Wu, 1999).

Two short years later, the world of the Internet campaign could not be more different. Few Web sites shy away from such interactive features today. Success stories such as those cited at the beginning of this chapter demonstrate the potential of the Internet as a tool for recruiting volunteers, controlling press coverage, and amassing a campaign war chest. More recent studies of campaign use of the Web demonstrate that the sites have become graphically rich and highly interactive, with significant issue content and an overwhelmingly positive tone (Greer & LaPointe, 2001). The "rational" campaign, today, is partially an Internet campaign.

Individualized Campaigns?

The ability to create an "enhanced" Web site is a double-edged sword. On one hand, building in audio and video

extensions increases stickiness and improves a page aesthetically. Likewise, added customizability allows site owners to tailor messages to a specific audience, be that audience political or commercial. With such advantages come tradeoffs, however, both in time and money and in heightened consumer expectations.

The incorporation of images, sound, and movies was described previously in emphasizing the democratization of new technologies for the purposes of discussion and debate. These same technologies tend to originate in the hands of those with a major Web presence: large and well-established interest groups, political parties, and their preferred candidates. Smaller interest groups, fringe political parties, and less well-funded political candidates have slowly followed suit. The same trickle-down trends have held for extensions of Web pages, such as message boards, chat rooms, and opinion polls. To enhance a site in these ways requires significant monetary investments for both site creation and maintenance. The sites are physically larger, consuming disk space, processor capacity, and bandwidth that previously were unneeded. Content creation requires yet more equipment as well as user training and the time inherent in recording, editing, and polishing. Finally, software packages for features such as message boards may be used "off the shelf," but typically customization is needed above and beyond installation (not to mention policing of posts and other clerical work). The payoff cited by Sadow and James in the commercial realm is tangible, but so are the expenses.

Customization is another dilemma altogether. As the drive to push campaigning online grows in coming years, candidates and interest groups will feel obligated to adjust their sites to the desires of each individual user (or at the very least each class of users). Business has already begun to deal with the pros and cons of customizing sites, and the experiences of such corporations are instructive for coming applications to the political sphere. J. Scott Sanchez, formerly employed by Procter & Gamble and now part of Intuit's Quicken team, notes that

> One of the long held goals of traditional marketers has been to send the right message to the right person, since every consumer tends to have a slightly different view of things. In the past, this was impossible and marketers just relied on mass advertising to try to get a consistent message to as many people as possible. However, with the advent of the Internet, it will now be possible to tailor messages to specific individuals. (Sanchez, 2000)

Replace the word "consumer" with "voter" and "marketers" with "campaign workers" and it yields an equally compelling message.

The promise of customization is one of the driving forces behind numerous online ventures, from Internet portals (My Yahoo) to music sites (My.MP3.Com). Often in registering for a custom site a user will provide the site owner with marketing information such as an e-mail address as well, adding to the allure. Although Sanchez notes that "the message is perfectly targeted and its effectiveness rockets upwards," he also points out that "one of the interesting repercussions of this individualized marketing, however, is that companies now may be held more accountable for their promises. Because individuals are receiving a tailored e-mail that promises to do a certain task in a certain way, such as 'gets the whites whiter,' consumers may feel betrayed if it does not." Again, a parallel exists in politics. Cass Sunstein, a legal scholar at the University of Chicago, worries about the customization of our Internet experience, because we are not forced to confront opinions and ideologies different from our own (Sunstein, 2002). Personalization of campaigns is problematic for campaigns as well. Although it does allow a custom message to be delivered to a potential voter, a politician or interest group opens the door to conflicting or at mutually nonsatisfiable promises. After all, one of the main reasons for political parties and elections is that people are forced to choose among "bundles" of less than ideal, yet feasible, alternatives. In the individualized world of the Internet, everyone might feel that government must satisfy his or her particular bundle of desires. The result, according to one observer, could be "accelerated pluralism," a further breaking down of coherent political communities (Bimber, 1998).

Thus, candidates have found the Internet to be a viable source for recruitment, campaign fundraising, and mobilizing voters. The Internet, then, may empower individuals, but only if they are the sort of individuals that candidates wish to reach. Furthermore, even if candidates, by using the Internet, motivate far more citizens to participate, the individualization of the Internet experience may result in an electorate that is more polarized and pluralized than at present.

Interest Groups and Political Parties on the Web

The Internet thus far has revolutionized commerce, as well as much of day-to-day social interaction. The capability of the Internet to act as a post office and an interactive, worldwide accessible bulletin board, as well as a real-time source of information, will likely impact the political arena in important ways. Beyond political candidates, the role of intermediary groups in politics is likely to be affected dramatically simply because the essence of the Internet lies in its potential to connect. Intermediary groups, organizations who act as the connecting tissue between the mass public and the governmental elite, are the political players most likely to benefit from the convenient tools for communication and organization that the Internet makes readily available.

The Internet lowers the cost of communication. There are a number of regular chores the Internet makes easier and faster. Because of the low transaction costs, some have claimed that the Internet will result in a more even playing field between interest groups with abundant resources and those with much less. Indeed, some have even gone so far as to say that the Web is "potentially the greatest thing since the postal system and the telephone for political groups" (Hill & Hughes, 1998, p. 133). Others however, have claimed that, although the Internet may make things cheaper overall, there are still prohibitive costs, and there, as everywhere else, resources still matter. Regardless, the

spread of the Internet has already affected the way that interest groups conduct their activities and will continue to do so in the future.

The importance of fundraising for any interest group is readily apparent; groups require financial support to continue operating. As Richard Davis notes, "[g]roups have formed in competition with each other, but they are not guaranteed equal voices or shares in power . . . (m)ost policy maker attention is centered on groups who possess substantial resources" (Davis, 1999, p. 81). Fundraising can take a variety of forms, especially with regard to groups dependent on businesses or other special parties for support. This fundraising carried out by interest groups is a type of direct mobilization, where political leaders communicate directly with citizens and provide an opportunity for political action. A request for members to volunteer time to support the organization is one common example of this. Another important way direct mobilization occurs is in the basic task of educating the public and the group's members along with informing them of news and events related to the group. This process is vital because an informed membership is more likely to care about the group's issues and actively support the group in some way.

Other types of direct mobilization include requests to sign petitions and write letters to political representatives. These efforts to encourage individuals to contact the government, described as "outside" lobbying by Ken Kollman or "grassroots" lobbying by Mark Petracca, constitute an important tactic for interest groups to use to achieve results. Kollman argues that this outside lobbying performs the dual tasks of "communicat[ing] aspects of public opinion to policymakers" and "influenc[ing] public opinion by changing how selected constituents consider and respond to policy issues" (Kollman 1988). Petracca (1992) emphasizes its widespread use, stating that "interest groups across the political spectrum now pursue grassroots lobbying with a vengeance." In this way, interest groups encourage direct contact between their members and government to further their own ends.

Because communication is so central to an interest group, this has the consequence of making its main cost the cost of communication. The traditional methods of mass media advertising, telephone campaigns, and mass direct mailings all incur significant costs to the group performing them. The potential of the Internet, then, becomes clear. The difference in cost between 1,000 and 100,000 people reading an informative Web site put up by an interest group is most likely trivial (due to bandwidth charges) or zero; however, the cost of printing and mailing 100,000 brochures is presumably much higher than that of doing so for only 1,000. Thus interest groups can reach a much larger audience without incurring higher transaction costs through use of the Internet.

A similar logic can be applied to member responses to group requests as well as member communication to a group or the government in general. Well-written form letters can be sent online with the mere click of a button. People would (generally) like to spend less time on the task and therefore prefer the easier online method. This can be extended to essentially all exchanges that take place between a member and a group: joining, donations and sales, getting current news and events, and providing feedback to the group.

The Internet also presents the opportunity for groups to make communication between and among members easier. Web forums and online services, such as electronic greeting cards, enable Web sites to develop a community made up of regular visitors to the site. Fronting the resources necessary for this effort can pay off for the interest group as well, because these new social networks will discourage members from quitting, encourage members to be active, and possibly even attract new members, as entrance to this online community becomes another type of solidary incentive (Olson, 1971).

In summary then, an interest group's or political party's success is affected significantly by three types of communication: group-to-member, member-to-group, and member-to-government. Also, the interest group can help itself by encouraging social networks among its members, or member-to-member interaction. The Internet has the potential to greatly decrease the transaction costs for all of these types of communication.

This suggests that interest groups should and will pursue online options for their activities. This capability of the Internet to decrease costs and provide alternative methods of communication is precisely what gives it huge relevance to politics. So, in theory, Internet usage is a valuable pursuit for interest groups in a variety of ways. However, the issue of efficiency is still largely ignored. The common thinking goes that, because Web site construction and Internet use are relatively cheap, then if such efforts produce any results, they must be worthwhile. With these low production costs, it should be expected that there would be roughly equivalent Web usage across interest groups with different budgets. Or if differences in breadth of group interests are considered, then there should be at least no direct correlation between a group's budget and its Web presence, as the whole concept is that the low cost enables any group to provide as large an online presence as it desires.

As with studies of campaign communications, however, there are few up-to-date studies of the efficacy of interest group and political party activities on the Internet(although see the studies conducted for the author by Tang and Looper [1999] and Casey, Link, and Malcolm [2000] available online at http://www.reed.edu/~gronkep/webofpolitics). It is clear that the Web sites are being created, but at what cost and for what impact? Can interest groups enhance democratic politics by substantially increasing political participation? Those few studies that have been conducted examine political party Web sites and conclude that established interests dominate this new medium as they did traditional avenues of political competition (Gibson & Ward, 1998; Margolis, Resnick, & Wolfe, 1999). No comparable studies of interest group sites have been conducted. For now, the question remains open.

The Hotline to Government? The Internet and Direct Democracy

Imagine that a federal agency such as the Environmental Protection Agency is holding a hearing on a new set of

regulations in the year 2010. Rather than simply scheduling a public comment session in Washington, anyone is allowed to register opinions via Internet teleconferencing. Local citizens, concerned politicians, and informed observers play on a level playing field with the moneyed interests and high-powered lobbyists who so often seem to dominate federal decision making.

Alternatively, imagine a world (already in place) where broadband could provide the electorate insider access to all levels of government. C-SPAN already provides gavel-to-gavel coverage of congressional debates and hearings. Audiovisual technologies might replace e-mails to a congressman, which usually receive an automatic response reply, with short question and answer sessions conducted live over Internet teleconferencing (with a congressional aide, if not with the congressman himself).

The most groundbreaking aspect of the Internet might be the ability of citizens to express their opinions *directly*, bypassing parties and interest groups. Political scientists have long realized that citizens vary in their preference for different "modes" of political participation. Some vote, others attend rallies, still others prefer to write letters. This is precisely what we would expect when individuals vary so much in their access to political resources and their integration into social networks (Rosenstone & Hansen, 1993). What difference might the Internet make? In a wired world, it is far easier (perhaps too easy) to dash off an e-mail to a member of Congress or offending bureaucrat. At the same time, just as members have had to contend with reams of postcards generated by grassroots lobbying efforts, universal access to e-mail is likely to reduce its impact. Ironically, then, the Internet and e-mail have made the old-fashioned handwritten and signed letter far more effective, simply by contrast.

Are these visions likely to become a reality? Due to the stipulations of the 1974 Administrative Procedures Act (A.P.A.), these agencies are obligated to hear from everyone who would like to speak on an issue facing the agency prior to the agency's ruling. The procedure further states the agency must take all arguments into account when rendering a ruling and provide reasons for its decision. Currently, speaking before an agency like the E.P.A. carries the high costs of a trip to Washington D.C. With the advent of e-mail and the World Wide Web, citizens can easily collect information and express their opinions on new regulations and public laws. Once the aforementioned technological enhancements become commonplace, it may even be possible for "teletestimony" to be given at congressional hearings and agency public comment sessions.

Some optimists, such as Andrew Shapiro, further predict that individualized *control* over the means and modes of contact with government will empower individuals (Shapiro, 1999). As Kevin Hill and John Hughes point out, the Internet's low costs have created tremendous new opportunities for fringe groups seeking to become more recognized. A fringe political group with limited resources can create a Web page that differs little in quality from a Web site for a well-financed political party (Hill & Hughes, 1998, p. 134). This logic also applies to fringe group participation in local, state, and federal political activities. Via Internet teleconferencing, a radical environmentalist group operating on a budget of $10,000 a year could afford to present its ideas before the E.P.A. in the same manner as the Sierra Club.

In addition to fringe groups, teleconferencing also aids those political activists confined by the costs of mobility. This group includes stay-at-home mothers and fathers, senior citizens who are unable to travel without assistance, and the disabled. Assuming that a person in one of these groups was politically motivated yet constrained by his or her situation, teleconferencing could mobilize that citizen by allowing him or her to participate. In theory, the mobilizing effects of ubiquitous high-speed Internet access and enhanced audio/visual capabilities could create an even more powerful lobbying force for organizations such as the A.A.R.P. or women's rights movements, assuming that these organizations are stripped of some influence by "immobile" members who might otherwise directly participate in lobbying, protesting, or debating before a governmental body.

CONCLUSION

E-commerce was supposed to revolutionize the business world, making "bricks" a thing of the past. The post-Internet hangover has demonstrated the importance of the "old rules" of investment and the preference among consumers for bricks over clicks. Similarly, the "old rules" of politics, the basic relationships between individual motivations, organizational effort, and political action, have remained stubbornly resistant to the lure of computer revolution. Picnics and pig pickin's, state fairs, and outdoor rallies remain an important part of the "retail center" of American politics. Most candidates still spend the vast bulk of their advertising dollars on traditional media outlets (television, radio, and newspapers) or direct mail contacts, rather then choosing to contact voters via the Internet. Most political parties continue to spend tens of millions of dollars each campaign cycle on traditional political activities, such as voter registration drives, political canvassing, and "get out the vote" efforts. Even so-called "high-tech lobbying" efforts (West & Loomis, 1998), although taking full advantage of electronic technologies in order to educate citizens and mobilize participants, continue to focus their efforts on traditional media outlets, grassroots organizing, and old-fashioned lobbying in the halls of the Capitol.

The Internet *has* become a central tool for mobilization efforts by political organizations, as the rational choice approach to voting would predict. The individual has little incentive to get involved politically, but organizations have great incentives to mobilize. The increase in electronic mail and Web access, the growth in broadband access, and the seeming inevitability of Internet commerce, has opened up a new frontier for both citizens and elites. Enhancements in audiovisual capabilities could lower the costs of participation for groups that previously could not overcome the high costs of transportation to Washington. Candidates could attract new political participation via "stickier" Web sites. The promise of online voting remains unproven, but given the rapid expanse of Internet access and computer ownership, online voting and referenda could mobilize previously underrepresented portions of the population.

Yet, although the "new political machine" holds the potential for a more democratized and decentralized political system, to date it has primarily reinforced preexisting biases in political participation and influence. The Internet has not changed significantly the way we have understood mass democracy for over 200 years (Bimber, 1999). Changes in the means of participation will constantly evolve to match the most current technology available; anticipating changes in the number and type of people who participate will continue to be an unpredictable science.

GLOSSARY

Collective action problem A situation where individuals choose not to work toward the provision of a public good because the costs to them individually exceed the benefits which they receive, so that no one participates in the provision of public goods.

Deliberative polling A survey polling technique promoted by James Fishkin, where poll respondents participate in an open discussion for a period of time before choosing options.

Grassroots lobbying Lobbying efforts that focus on stimulating activities by citizens, such as formation of local groups, letter-writing, and e-mailing.

Individualization The ability of an Internet user to individualize or personalize his or her news-gathering experience. Also referred to as "customization" and "personalization."

Intermediary organizations Organizations, such as political parties or interest groups, that stand in between the mass public and government.

Mass public Contrasted with elites, the mass public comprises the vast bulk of the population.

Mobilization Efforts by organizations and individuals to stimulate and encourage political involvement and participation.

Modes of political participation The varied ways that citizens may choose to influence government, including campaigning, writing letters, joining groups, and protesting.

Netizens Term used to describe "citizens" of an Internet community.

Political elites Contrasted with the mass public, elites are that segment of the population that is better informed, educated, and interested in politics. Sometimes used to describe decision makers.

Political machine Tightly organized political organizations that tend to exchange benefits (jobs, social welfare) for votes; existed in many American urban areas in the early 20th century. Also sometimes described as "boss politics."

Public good A good such that if it is provided to anyone in a group, it must be provided to everyone in a group (e.g., national defense, clean air). Public goods often suffer from the collective action problem.

Rational choice Theory of individual action that assumes goal-seeking behavior, while maximizing benefits and minimizing costs.

Rational ignorance Assumption that some individuals will choose to ignore political events, news, and the like because the costs of being informed exceed the benefits from such information.

Referendum Election format where voters choose among a set of legislative options; also described as "direct democracy" and "initiative government." Common in the Western United States.

Solidary incentive The feelings of belonging and community that accrue to those who join a group working to provide public goods.

Stickiness Characteristic of a Web site that encourages viewers to remain on that site.

Strong talk Theory of democracy promoted by Benjamin Barber that encourages high levels of citizen discussion, deliberation, and participation.

Social capital The web of social and personal relationships that encourage participation in community and civic affairs.

Town meetings Form of political decision making where the members of a community gather together, discuss options, and vote on alternatives.

Virtual community Contrasted with physical communities, which are defined by geographic space, virtual communities exist in virtual or cyberspace.

CROSS REFERENCES

See *Developing Nations; Digital Divide; Electronic Commerce and Electronic Business; Internet Etiquette (Netiquette); Internet Literacy; Legal, Social and Ethical Issues; Online Communities*.

REFERENCES

Aldrich, J. A. (1993). Turnout and rational choice. *American Journal of Political Science, 37*(1), 246–278.

Ariely, D. (1998). *Controlling the information flow: On the role of interactivity in consumers' decision-making and preferences*. Ph.D. dissertation, Duke University.

Barber, B. R. (1984). *Strong democracy: Participatory politics for a new age.* Berkeley: University of California Press.

Bimber, B. (1998). The Internet and political transformation: Populism, community, and accelerated pluralism. *Polity, 31*(1), 133–160.

Bimber, B. (1999). Information and the evolution of representative democracy in America: From *The Federalist* to the Internet. Unpublished manuscript, Department of Political Science, University of California, Santa Barbara.

Bimber, B. (2002). *Information and American democracy.* New York: Cambridge University Press.

Chiu, L. (2000, March 25). Record primary turnout; Dem's vote attracted across racial lines. *The Arizona Republic*, p. B1.

Corrado, A. (2000). *Campaigning in cyberspace.* Washington, DC: The Aspen Institute.

Davis, R. (1999). *The web of politics*. New York: Oxford University Press.

Fishkin, J. (1991). *Democracy and deliberation*. New Haven, CT: Yale University Press.

Gibson, R. K., and Ward, S. (1998). U.K. political parties and the Internet: "Politics as usual" in the new media?

Harvard International Journal of Press Politics, 3(3), 14–38.

Graber, D. (1984). *Processing the news.* New York: Longman.

Graber, D. (2001). *Processing politics: Learning from television in the Internet age.* Chicago: University of Chicago Press.

Green, D. P., & Shapiro, I. (1996). *Pathologies of rational choice theory.* New Haven, CT: Yale University Press.

Greer, J., & LaPointe, M. E. (2001). *Cyber-campaigning grows up: A comparative content analysis of senatorial and gubernatorial candidates' web sites, 1998–2000.* Paper presented at the Annual Meeting of the American Political Science Association.

Hansen, J. M. (2001). *To assure pride and confidence in the electoral process.* Final report from the National Commission on Election Reform. Retrieved August 15, 2002, from http://www.reformelections.org

Hennessy, J., & Patterson, D. (1990). *Computer architecture: A quantitative approach.* San Francisco: Morgan Kaufmann.

Hamilton, A., Madison, J., and Jay, J. (1961). *The Federalist papers* (C. Rossiter, Ed). New York: New American Library.

Hill, K., & Hughes, J. (1998). *Cyberpolitics*. Lanham, MD: Rowman & Littlefield.

Holtz, D. (2000, February 22). Berkeley residents can take action on Internet. *San Francisco Chronicle,* p. A13

Horrigan, J. B., & Rainie, L. (2002). *The broadband difference: How online Americans' behavior changes with high-speed Internet connections qt home.* Report issued by the Pew Internet and American Life Project. Retrieved August 17, 2002, from http://www.pewinternet.org

Kamarck, E. C., & Nye, J. (1999). *Democracy.com: Governance in a networked world.* Hollis, NH: Hollis Publishing.

Lessig, L. (1999). *Code and other laws of cyberspace.* New York: Basic Books.

Lessig, L. (2002). *The future of ideas.* New York: Random House.

Margolis, M., Resnick, D., and Wolfe, J. (1999). Party competition on the Internet in the United States and Britain. *Harvard International Journal of Press Politics, 4*(4), 24–47.

Moore, G. E. (1965). Cramming more components onto integrated circuits. *Electronics, 38*(8). Retrieved August 10, 2002, from ftp://download.intel.com/research/silicon/moorespaper.pdf

Norris, P. (2001). *Digital divide: civic engagement, information poverty, and the Internet worldwide.* New York: Cambridge University Press.

NUA Internet Surveys (2002, February). *How many online?* Retrieved August 15, 2002, from http://www.nua.ie/surveys/how_many_online

Olson, M. (1971). *The logic of collective action.* Cambridge, MA: Harvard University Press.

Pew Research Center for the People and the Press (1996). *News attracts most Internet users.* Washington, DC. Retrieved August 18, 2002, from http://www.people-press.org

Pew Research Center for the People and the Press (2002, June 9). *Public news habits little changes by September 11*. Retrieved August 18, 2002, from http://www.people-press.org

Phillips, D. (1999). *Are we ready for Internet voting?* Report from the Voting Integrity Project. Retrieved January 20, 2002, from http://www.voting-integrity.org

Phillips, D. (2000). *Is Internet voting fair?* Report from the Voting Integrity Project. Retrieved January 20, 2002, from http://www.voting-integrity.org

Podesta, J. A. (2002, May/June). Is the Internet a hopeless model? *Ideas Magazine*.

Putnam, R. D. (2000). *Bowling alone.* New York: Simon and Schuster.

Rosenstone, S., and Hansen, J. (1993). *Mobilization, participation, and democracy in America.* New York: Macmillan Publishing.

Sadow, J., and James, K. (1999). *Virtual billboards? Candidates web sites and campaigning in 1998.* Paper presented at the Annual Meeting of the American Political Science Association.

Sanchez, J. S. (2000, April). Telephone interview with the author conducted by Brian Stempel, student in "Politics and the Internet" course at Duke University.

Sarkar, D. (2000, December 4). Web an election winner. *Federal Computer Week*. Retrieved August 18, 2002, from http://www.fcw.com/civic/articles/2000/dec/civ-comm1-12-00.asp

Shapiro, A. (1999). *The control revolution: How the Internet is putting individuals in charge and changing the world we know.* New York: Public Affairs Press.

Sunstein, C. (2002). *Republic.com.* Princeton, NJ: Princeton University Press.

Thornburg, R. (2001). *Digital donors: How campaigns are using the Internet to raise money and now it's affecting democracy.* Paper #1, Occasional paper series, Democracy Online Project. Washington DC: George Washington University.

Turkle, S. (1997). *Identity in the age of the Internet*. New York: Touchstone.

Verba, S., Schlozman, K. L., and Brady, H. (1995). *Voice and equality.* Cambridge, MA: Harvard University Press.

West, D., and Loomis, B. (1998). *The sound of money.* New York: Norton.

Wu, G. (1999). Perceived interactivity and attitude toward Web sites. In M. Roberts (Ed.), *Proceedings of the 1999 Conference of the American Academy of Advertising* (pp. 254–262). Gainesville, FL: University of Florida.

Privacy Law

Ray Everett-Church, *ePrivacy Group, Inc.*

INTRODUCTION

Understanding privacy is a true challenge, in no small part because of the difficulty in defining the concept of privacy itself. The textbook definition of privacy only begins to scratch the surface of a deeply complex issue, made all the more complex because of the strong personal feelings evoked by privacy breaches. Accounting for privacy concerns can be a daunting task, especially when one is building Internet-based services and technologies for which success can depend on not offending consumers' mercurial sensibilities about the value of their privacy versus the value of those services that depend on free-flowing personal data.

This chapter discusses the roots of privacy law, including the various ways that privacy matters are dealt with under constitutional law, statutes, and common law. With the fundamentals established, the rest of this chapter discusses how many of those principles have come to be applied in today's Internet-oriented privacy terrain and how businesses must prepare for doing business in this new environment.

PRIVACY LAW BASICS

Privacy Defined

The *Merriam-Webster Dictionary of Law* defines privacy as "freedom from unauthorized intrusion: state of being let alone and able to keep certain especially personal matters to oneself." Within this broad "state of being let alone," particular types of privacy intrusion have been recognized under law. How one defends oneself against intrusions differs, however, based on who is doing the intruding.

Constitutional Privacy

Even though one will find no trace of the word "privacy" in the U.S. Constitution, a series of Supreme Court decisions beginning in the 1920s began to identify the modern concept of privacy. As the court refined its views on the subject, it found the idea of privacy within the spirit of the Constitution's protections, if not in the plain language of the document. In 1928, in a landmark wiretapping case (*Olmstead v. United States,* 1928), Supreme Court Justice Louis Brandeis articulated the following ideas in some of the most important words ever written about privacy:

> The makers of our Constitution undertook to secure conditions favorable to the pursuit of happiness. They recognized the significance of man's spiritual nature, of his feelings and of his intellect. They knew that only a part of the pain, pleasure and satisfactions of life are to be found in material things. They sought to protect Americans in their beliefs, their thoughts, their emotions, and their sensations. They conferred, as against the Government, the right to be let alone — the most comprehensive of rights and the right most valued by civilized men. (Brandeis dissenting, *Olmstead* at 478)

Brandeis's phrase "the right to be let alone" is one of the most often-repeated ideas in privacy and has influenced the court's inquiry beyond the plain words of the Bill of Rights to find other privacy rights that are logical extensions of the meaning contained in the original words, including the following:

- The First Amendment right of free speech has been read to include the right to speak anonymously. Free speech has also been interpreted in reverse: You have the right to not be forced to say certain things.
- The First Amendment right of free association means that you can join clubs and affiliate yourself with anyone you choose. Inherent in that right, according to the

court, is the right not to say with whom you're associating.
- The Fourth Amendment prohibits the government from searching your home and property and from seizing your papers or possessions, except under very specific circumstances. The Fourth Amendment has also been read to give certain rights against government wiretaps and surveillance.
- The Fifth Amendment includes various rights of due process, which means that if the government is interested in depriving you of any of your rights—throwing you in jail, for example—it must first follow strict procedures designed to protect your rights. Among those is the right against being forced to incriminate yourself.

The equal protection clause of the Fourteenth Amendment requires that both sexes, all races, and all religions be given equal protection under all the laws of the United States and all the laws of every state. This protection comes despite other amendments that can be read to permit some types of discrimination. These rights aren't absolute, however. For example, consider the following:

- The government can set up wiretaps, perform surveillance, and perform searches and seizures if it has reasonable belief ("probable cause") that a crime has been committed and if given permission (a "warrant") by a judge.
- The government can establish secret wiretaps and surreptitiously search your home or car, without a normal warrant, if you are suspected of being a terrorist or an "agent of a foreign power."
- Certain sexual activities, even between consenting adults in the privacy of their bedroom, can be illegal.
- It can be illegal to keep certain materials in your home, such as drugs or child pornography.
- Certain public organizations (such as the Jaycees, which was the subject of a lawsuit that established this precedent) cannot use the First Amendment right of free association to exclude protected classes of people, such as women or certain minorities. On the other hand, at the time this book was written, the Boy Scouts could discriminate against gay people.

But the Constitution only affects privacy issues involving the government. What are your rights against people who are not part of the government, such as individuals and corporations? That's where a patchwork of common-law privacy protections and several statutes comes into play.

Common-Law Privacy

The common law is a set of rights and obligations first recognized by courts rather than by legislatures. Just because it is "judge-made" law, however, one cannot discount the common law as being less forceful. In fact, many common-law rights have been enforced for centuries and are some of the most powerful precedents in our legal system. They are rarely overturned by legislatures, and many state and federal laws are simply codifications of common-law ideas that have been around for hundreds of years.

In a groundbreaking law review article in 1960, William Prosser set out four broad categories of common law that underlie privacy-related torts:

- Intrusion into one's seclusion,
- Disclosure of private facts,
- Publicizing information that unreasonably places one in a false light, and
- Appropriation of one's name or likeness.

Intrusion

The tort of intrusion recognizes the value of having your own private space and provides relief from those who would seek to violate it. Eavesdroppers and "peeping toms" are two examples of activities considered intrusion.

Disclosure

The tort of disclosure recognizes that making public certain private facts can cause harm to an individual. For example, disclosures about someone's health status, financial records, personal correspondence, and other kinds of sensitive personal information can cause harm if made public.

False Light

The tort of false light is similar to libel in that it involves publicizing falsehoods about someone, but it is subtly different. One famous case of false light, *Cantrell v. Forest City Publishing Co.* (1974), involved a family who was inaccurately portrayed in a news article in a humiliating fashion that brought shame and embarrassment. Another, *Douglass v. Hustler Magazine* (1985), involved a model who posed nude for a popular pornographic magazine, which were instead published with embarrassing captions by a notoriously vulgar magazine instead.

Appropriation

This tort involves using the name or likeness of someone for an unauthorized purpose, such as claiming a commercial endorsement by publishing someone's image (or even that of a look-alike impersonator) in an advertisement.

In this age of modern technology, there appear to be many new ways of violating these centuries-old privacy torts. The prevalence of miniature "Web-cams," highly sophisticated digital photo editing applications, and the vigorous online trade in pornographic imagery, have each added to the ways in which individual privacy can be violated.

PRIVACY LAWS IN THE UNITED STATES AND ABROAD

In a 1973 report to Congress, the U.S. Department of Health, Education and Welfare (HEW) outlined four tenets of fair information practices. These guidelines were groundbreaking in that they set forth four characteristics that any fair policy regarding the collection and use of

personal information had to take into account. The four tenets were as follows:

1. *Notice*. Details of information practices and policies should be disclosed to data subjects.
2. *Choice*. Data subjects should be given the ability to exercise choices about how data may be used or disclosed.
3. *Access*. Data subjects should be permitted access to data gathered and stored about them.
4. *Security*. Holders of personal data should be responsible for providing reasonable levels of security protection for data in their possession (HEW, 1973).

Since then, there have been a number of laws enacted in the United States dealing with individual privacy. The standard U.S. approach is, however, to focus on particular types of information used by or about specific sectors:

- *Banking records.* Your personal banking information is protected by law, up to a point, including under provisions of a new law called the Financial Services Modernization Act (also known by its authors as the Gramm–Leach–Bliley Act).
- *Credit reports.* The Fair Credit Reporting Act (FCRA) require that credit bureaus handle your data in certain ways.
- *Medical and Health Insurance Records.* Laws and regulations governing how medical records can be used have been in place for several decades, and provisions of a new law called the Health Insurance Portability and Accountability Act (HIPAA) are creating new rights for patients to protect and access their own health information (U.S. Department of Health and Human Services, 2002).
- *Government records.* The Privacy Act of 1974, which included the original tenets outlined in the HEW report, sets limits on how government agencies can collect and use personal information, whereas laws like the Freedom of Information Act of 1966 require government to give all citizens access to certain government records, provided that the government also take precautions not to breach privacy when making that information public.
- *Children's Privacy.* Although not limited to one business sector, a law called the Children's Online Privacy Protection Act of 1998 (COPPA) places restrictions on online organizations that seek to collect data from one sector of the public: children under the age of 13. COPPA requires the publication of a privacy policy to explain data practices relating to children's information, requires verifiable parental consent before any personally identifiable information may be collected from children over the Internet, and limits companies ability to share children's information with third parties.

International Privacy Law

The recognition of privacy rights in international law goes back to December 10, 1948, when the United Nations (UN) adopted the Universal Declaration of Human Rights. Article 12 of that document says, "No one shall be subjected to arbitrary interference with his privacy, family, home or correspondence, nor to attacks upon his honour and reputation. Everyone has the right to the protection of the law against such interference or attacks" (UN, 1948).

Building on that foundation and applying the four tenets articulated in 1973 by the U.S. government, in 1980 the multinational Organization for Economic Cooperation and Development (OECD), of which the United States is a member, issued its eight Principles of Fair Information Practices. These principles consisted of the following:

- *Collection Limitation.* There should be limits to the collection of personal data, and any such data should be obtained by lawful and fair means and, where appropriate, with the knowledge or consent of the data subject.
- *Data Quality.* Collection of personal data should be relevant to the purposes for which they are to be used and, to the extent necessary for those purposes, should be accurate, complete and kept up-to-date.
- *Purpose Specification.* The purposes for which personal data are collected should be specified not later than at the time of data collection and the subsequent use limited to the fulfillment of those purposes or such others as are not incompatible with those purposes and as are specified on each occasion of change of purpose.
- *Use Limitation.* Personal data should not be disclosed made available or otherwise used for purposes other than those specified in accordance with principle of purpose specification, unless done with the consent of the data subject or by authority of law.
- *Security Safeguards.* Personal data should be protected by reasonable security safeguards against such risks as loss or unauthorized access, destruction, use, modification or disclosure of data.
- *Openness.* There should be a general policy of openness about developments, practices, and policies with respect to personal data. Means should be readily available of establishing the existence and nature of personal data, and the main purposes of their use, as well as the identity and usual residence of the data controller.
- *Individual Participation.* An individual should have the right to obtain from a data controller confirmation of whether data is held about the individual, to be given access to the data in an intelligible form, and to have the data erased, rectified, completed or amended.
- *Accountability.* A data controller should be accountable for complying with measures that give effect to the principles. (OECD, 1980)

The European Union has taken the OECD principles and incorporated them into a sweeping Data Privacy Directive that establishes these principles in law. The directive mandates the following minimum standards in all countries that are members of the European Union (EU):

- Companies can only collect information needed to complete the transaction, and must delete it after the transaction is over, unless they have explicit permission.

- Consumer's personal information must be kept up to date, or deleted.
- The purpose for collecting data must be given at the time that data is collected.
- An individual's personal information cannot be used for any other purpose (such as mailing catalogs or coupons) unless a company has explicit permission.
- Companies must have appropriate security safeguards in place to guarantee privacy of any data in their possession.
- Companies must keep consumers advised in a clear and open manner about their data practices and how consumer's privacy will be impacted by any changes.
- Consumers must be permitted to see any information a company has on file about them, must be permitted to correct any errors, and must be allowed to delete data unless there's a legally mandated reason for keeping it.
- Companies who keep consumer information must have someone in the company accountable for ensuring that the privacy laws are being adhered to.
- Companies may not transfer data outside of the EU unless the country to which the data is being transferred has privacy laws as strict as those in the EU (European Commission, 1995)

It should also be noted that these restrictions apply to all data in a company's possession, whether customer data or employee data. And these are minimum standards; individual member countries can—and have—enacted laws that are even stricter. To enforce their privacy laws, many EU member countries have established data protection authorities—government agencies whose mandate is the policing of data practices within, and crossing, national borders. These authorities often require corporations who possess personally identifiable information about any citizen of their nation to register with the agency and file detailed statements of what data is collected and how it is used.

In addition, whereas U.S. law focuses on certain categories of information, such as financial or healthcare data, holders of the data such as credit bureaus, or categories of data subjects such as children, the EU law gives special consideration to data about

- Race,
- Religious affiliation,
- Membership in political parties and trade unions, and
- Criminal records.

These topics are of particular concern to Europeans, in part because of how records containing information about race, religion, and trade union memberships were gathered and used by the Nazi regime in Germany and in its occupied countries to decide who should be shipped off to concentration camps. For Europeans, the threat of private information being misused is more than a test of wills between marketers and consumers, but has meant the difference between life and death for the parents and grandparents of today's European lawmakers.

Cross-Border Data Flow

The issue of cross-border data flow has been particularly vexing for U.S. corporations, especially given the number of Internet-based firms with operations in the European Union that depend upon data flows from the EU back to the United States. Because the United States does not have broad privacy-protecting statutes on par with the EU, U.S. corporations face the prospect of being unable to communicate customer data, or even personnel records, back to U.S.-based facilities.

Recognizing the potential for numerous disputes, the United States and EU entered into a series of negotiations in the late 1999 and 2000, culminating in an agreement to create a Safe Harbor program. This program permits U.S. corporations to assert their adherence to an array of basic privacy requirements, with the assumption that those who certify compliance and bind themselves to enforcement measures in the event of misbehavior will be permitted to continue transferring data from the European Union into the United States (DOC, 2000).

BALANCING PRIVACY AND LAW ENFORCEMENT

In post-September 11 America, a great deal of public concern centers around the extent to which new antiterrorism intelligence-gathering will negatively affect the privacy of average citizens. Although few individuals will ever believe they merit the kind of surveillance activities implemented for mafia dons, drug kingpins, or terrorists, many are concerned that ubiquitous surveillance capabilities will result in less privacy for everyone, average citizens and mafia dons alike. Therefore, it is appropriate to discuss briefly the kinds of issues raised by increasing surveillance capabilities and to discuss a number of programs and laws that are adding to the pressures on personal privacy. More significantly, given the extent to which American business is increasingly becoming the repository of detailed information about the lives and business transactions of individuals, it is also appropriate to discuss how businesses are increasingly being called upon to aid law enforcement in their investigatory efforts, and why businesses need to exercise some judgment in deciding when and how to comply with law enforcement requests.

Surveillance, searches and wiretaps raise extremely complex legal and technical issues that are impossible to cover in this brief space. Should these issues arise in your personal or professional activities, it will not be possible for you to deal with them without the assistance of qualified legal counsel. There are, however, some things to keep in mind that will help you to understand how an organization may be affected.

Most domestic wiretapping is governed by the Electronic Communications Privacy Act of 1986 (ECPA). In addition, the Foreign Intelligence Surveillance Act of 1978 (FISA) governs wiretaps and surveillance of those considered "agents of a foreign power." Both ECPA and FISA were modified, clarified, and in some cases expanded significantly, by the Uniting and Strengthening America by Providing Appropriate Tools Required to Intercept and

Obstruct Terrorism Act of 2001, or USA PATRIOT Act for short.

ECPA

ECPA generally prohibits providers of communications services (e.g., Internet service providers) from disclosing the contents of an electronic communication, whether it is in transmission or in storage, to any person other than the intended recipient. ECPA also contains a number of exceptions, however, some of which include the following:

Service providers may make disclosures to law enforcement if proper warrants are presented. ECPA explains those procedures in some detail.

ECPA's limitations only apply to services offered to the public, not to operators of, for example, an internal corporate system.

ECPA does not restrict the collection, use, or disclosure to nongovernmental entities, of transactional information such as email addressing and billing information.

Disclosures to private parties pursuant to subpoenas issued by civil courts may also be permitted.

In addition, ECPA permits the government to request "dialing and signaling" information from telephone companies. Under these so-called "trap and trace" orders, law enforcement can use devices known as "pen registers" to capture the numbers being called and other information about the communications, short of the actual contents of the calls themselves. The contents of the calls can also be gathered, but only under a separate warrant that requires much more rigorous procedures and additional judicial review.

FISA

In cases in which information is sought about the activities of agents of foreign powers, such as terrorists or spies, law enforcement may seek disclosure of information relevant to an investigation through a special warrant procedure. There are two noteworthy differences between standard warrants and FISA warrants: First, FISA creates a system of special "FISA courts" in which judges meet, hear evidence, and issue warrants in total secrecy. Second, FISA warrants are much more sweeping than normal warrants and are not required to meet the same evidentiary standards as normal warrants. These differences raise significant Constitutional questions that have been raised in recent challenges to the activities of the FISA courts. Ironically, the FISA courts themselves have not been oblivious to the questions their seemingly unchecked powers have raised: A recently released decision of the FISA appeals court—the first document ever released publicly by the body—cited dozens of cases in which law enforcement provided deceptive or outright false information to the court in support of wiretap applications. Appealing to the U.S. Supreme Court, the Bush administration successfully overrode the FISA appeals court's objections to expanded wiretap procedures (EPIC FISA Archive, 2003).

Concerns about state-sponsored collection of data about individuals are nothing new. Privacy watchdogs and investigative journalists have widely publicized programs such as the FBI's "Carnivore" (a device for intercepting and recording Internet-based communications; EPIC Carnivore Archive, 2001), "Magic Lantern" (a piece of software that can be surreptitiously installed on a targeted computer, allowing law enforcement to capture every keystroke; Sullivan, 2001), and the rumored international wiretapping consortium called "Echelon" (EU Parliament, 2001).

Most recently, the U.S. Department of Defense sought funding of an antiterrorism program called "Total Information Awareness" which would have compiled electronic records on nearly every business, commercial, and financial transaction of every U.S. citizen. The massive database would then be analyzed in an effort to uncover transactions and patterns of behavior that could be deemed suspicious. Although the Total Information Awareness program was stripped of most of its funding by Congress in early 2003, the Department of Defense has vowed to keep researching the issues and technologies needed to undertake such a program (EPIC Total Information Awareness Archive, 2003).

Business Issues Under Wiretap Laws

The wiretap activities under ECPA and FISA have until recently been relatively limited in their effects on businesses. Aside from telephone companies and some Internet service providers, few businesses were affected by these procedures. Under recent changes to FISA made by the USA PATRIOT Act, however, law enforcement is now permitted to request business records from nearly any business to assist it in foreign intelligence and international terrorism investigations.

Previously, FISA only allowed law enforcement to request business records from certain categories of businesses, such as common carriers, hotels, and car rental facilities. Under the new rules, subpoenas can be issued without limit to particular categories, including banks, retailers, and any other entity within the government's reach. The USA PATRIOT Act also expanded the search and seizure from merely "records" to "any tangible things," such as computer servers.

The pen register and trap–trace provisions of ECPA have been expanded under the USA PATRIOT Act to add "routing" and "addressing" to the phrase "dialing and signaling," making it clear that these activities now include Internet traffic, not just telephone calls. The act does specify that the information retrieved through this process "shall not include the contents of any communication." There will undoubtedly be significant litigation in coming years to define where the dividing line falls between "content" and "addressing." For example, entering a search term or phrase into a search engine may cause the content of that search to be embedded in the address of the Web page on which the results are displayed.

PRIVACY ISSUES FOR BUSINESSES

In a widely published 2000 survey of more than 2,000 U.S. corporations, the American Management Association (AMA) discovered that 54% of companies monitor their employees' use of the Internet, and 38% monitor their employees' e-mail. In a follow-up survey in 2001, the percentage of companies doing Internet monitoring

rose to 63%, with 47% monitoring e-mail (AMA, 2001).

The rise in monitoring tracks with the rise in potential problems that can flow from providing access to the Internet. Along with the ability to work more efficiently, companies are now finding themselves held responsible when bad things find their way onto employees' desktops. In the same AMA study, 15% of the companies surveyed have been involved in some kind of legal action concerning employee use of e-mail or Internet connections. In several noteworthy cases, companies have been held liable for sexual harassment-related claims from harassment occurring over employer-operated message boards, employees leaving pornographic images on computer monitors, employees distributing sexually explicit jokes through office e-mail.

In response to these concerns, many companies have installed filtering mechanisms on their e-mail traffic looking for unacceptable language. Other companies have implemented software that blocks pornographic Web sites. Still others have opted for the low-tech approach of implementing zero-tolerance policies regarding the use of office computers for anything inappropriate.

Unfortunately, in some instances, these measures have resulted in confusion or wound up creating problems for both innocent and not-so-innocent people. For example, it was widely reported in 1999 that 23 employees of the *New York Times* were fired for trading dirty jokes over the office e-mail system (Oakes, 1999). Yet in other cases, recipients of unsolicited e-mail have opened the fraudulently labeled mail and been subjected to a barrage of pornographic images and salacious Web pop-up ads (Levine, Everett-Church, & Stebben, 2002).

Because Web monitoring logs and filtering systems may not be able to differentiate between Web pages viewed accidentally and those viewed purposefully, innocent workers can (and have) been left fearing for their jobs. For these reasons, companies are beginning to adopt internal privacy policies that help set better guidelines and establish reliable procedures for dealing with trouble when it arises.

Employee Privacy Policies

In most circumstances, there are few legal restrictions on what employers can do with their own computers and networks, up to and including monitoring of employee's communications. Although some firms quietly implement employee monitoring policies and wait to catch unsuspecting employees in unauthorized activities, many firms give notice to their employees that they may be monitored. Still others require employees to relinquish any claims of privacy as a condition of employment.

Increasingly, however, companies are recognizing the negative impact of paternalistic monitoring practices on employee morale. So to engender trust rather than inspire fear, increasing numbers of firms have begun providing their employees with privacy statements in their corporate employee handbooks or by publishing policy statements on internal Web sites. According to the AMA's 2001 survey, four out of five respondent firms have a written policy for e-mail use, and 77% for Internet use, 24% have training programs to teach these policies to employees, and an additional 10% plan one (AMA, 2001).

As noted earlier with regard to the European Union's Data Privacy Directive, companies with operations in the EU are already familiar with the mandate to provide data subjects—in these cases, employees—with information about the company's data-gathering and usage policies. Although there is currently no U.S. equivalent to these requirements, a growing number of firms are proactively recognizing that a well-defined set of privacy policies and practices can avoid misunderstandings and can even provide the basis of a legal defense in cases where companies are accused of failing to act on claims of Internet-based sexual harassment.

Developing an Employee Privacy Policy

The creation of a privacy policy for internal use in an organization can be as simple or as complex as the organization itself. Most companies collect information from their employees in the form of personnel records. Firms may also collect personal information from customers or clients. An internal privacy policy should address acceptable practices with regard to each type of information maintained by the company.

A good internal privacy policy should define what standards of behavior are expected of those who have responsibility over the data held by the company—including both employee data and the personal data of a company's customers—and should inform employees about the consequences of noncompliance. Additional topics that can be covered in a privacy policy include procedures for reporting breaches, procedures for allowing employees to access and correct their own personnel records, procedures regarding access to proprietary records such as customer lists, and procedures for auditing compliance and for training employees how to comply with the company's guidelines.

CONSUMER INTERNET PRIVACY

Before the Web existed, companies gathered whatever information they could get about their customers from a variety of sources, such as real estate transaction records, credit bureaus, court documents, and motor vehicle records. For many companies, among the most elusive, and hence the most valuable information—what you are interested in buying and exactly when you are ready to buy—was largely unavailable. Occasionally a clever marketer could devise an algorithm or a statistical model that might be used to infer some purchase preference from the tidbits of information that might be gathered about a customer from scattered sources. The Internet has made such information gathering much more commonplace.

Browser Privacy Issues

Many of the average computer user's online activities revolve around the two most popular Web browsers, Internet Explorer and Netscape. Browsers continue to evolve and improve, especially where privacy and security issues are involved. Even the most recent versions have some fundamental privacy problems that arise not by accident

but by design, however. In many cases, there are default settings that permit the collection and storage of usage data. These include the following:

- Browsers regularly tell Web sites what kind and what version of browser is being run, the operating system it is running on, and even what Web site "referred" the user to the current page.
- Some browsers have settings that permit users to capture and enter user IDs and passwords automatically for Web sites, as well as other personal information such as credit card numbers. These "wallet" features provide convenience but also present a privacy risk should anyone gain access to that machine and use it to log into sites or access users' personal information.
- Browsers can be instructed by Web sites to store little text files, called cookies, on local hard drives. Cookies can be used to store personal information or to assign unique identifiers that allow sites to identify users individually on future visits.
- Browsers can keep a log of every Web site a user visits and may even keep copies of the pages and images the user has viewed. The "history" function can log this data for days, weeks, or even months. Depending on the size of the hard drive and the default settings for a browser, it may also store days or weeks of Web page files and images in a "cache" folder.

Internet Explorer and Netscape have their own built-in privacy settings and controls. They vary in the level of control they allow over elements such as cookies, however. The "help" file that comes with each browser explains the browser's privacy settings and describes how to control them.

IP Addresses and Browser Data

In 1990, an engineer at a Swiss physics laboratory, Tim Berners-Lee, invented a new data-exchange standard in an effort to speed the sharing of information between researchers at widely dispersed locations. His creation was the hypertext transport protocol, or HTTP, and it made data sharing across the Internet literally as easy as point-and-click (Cailliau, 1995).

When the first Web servers and Web browsers were developed, however, not much attention was paid to subjects such as security and privacy. Because Berners-Lee and other engineers needed to troubleshoot their fledgling Internet connections, they built many automatic reporting features that would let them easily get to the root of the problem when something went haywire. This need for information such as browser type, version, operating system, and referring page was built into the earliest browsers and persists today.

Although not a tremendous privacy concern, the collection of this browser data is a standard function of most Web server software. Most sites collect this data for troubleshooting purposes and then delete it after some period of time, mostly because it can become very voluminous very quickly and its usefulness diminishes over time.

One element of the data that is also captured in the process of requesting and serving Web pages is the IP (Internet protocol) address of the user's computer. An IP address is a formatted string of numbers that uniquely identifies a user's computer out of all of the other computers connected to the Internet. IP addresses, which look something like 192.168.134.25, are assigned in blocks to Internet service providers, who in turn dole them out to their customers. With most dial-up Internet access accounts, users are assigned a "dynamic" IP address, meaning that the IP address assigned to a computer changes every time the user log onto his or her ISP, and gets tossed back into the ISP's pool of addresses when the user disconnects. By contrast, dedicated servers and some desktop computers in corporate or academic settings may have a "static" IP address, which is unique to that machine and may persist for the life of the equipment.

In this age of always-on Internet connections, however, such as those provided by DSL (digital subscriber lines) or cable modem services, it is possible for an average user's computer to have the same IP address for days, weeks, or months on end. From a privacy perspective, a static IP address can compromise one's privacy because an unchanging IP address make it easier for the truly determined to track an individual's Internet usage. For example, a site that collects IP addresses in its server logs may be able to correlate with other transactional records (e.g., purchase history or search parameters) to associate a unique IP address with a unique user and his or her online activities.

Given that most consumers use Internet service providers that regularly use dynamic IP addressing (as most of the DSL and cable modem providers claim), IP addresses are not considered a reliable means of allowing Web sites or online advertisers to track users uniquely. This lack of reliability should not be confused with anonymity. As a routine bookkeeping matter, many service providers log which IP address was allocated to which user's account at a given period of time. These connection records are frequently sought by prosecutors investigating criminal activities perpetrated via the Internet and by parties in private lawsuits over online activities. In recent years, dozens of companies have successfully uncovered the identities of "anonymous" critics by obtaining court orders for the release of user identities. Not every Internet service provider has willingly provided that information; in 2002, Verizon Internet fought attempts by the Recording Industry Association of America to release records identifying users accused of illegally trading music files. As of this writing, the federal district court in Washington, DC, held that Verizon was required to reveal the user's identity; however Verizon has appealed (McCullagh, 2003).

Cookies

Connections made using HTTP are called "stateless," which means that after the user's computer receives the content of a requested page, the connection between the computer and the faraway Web server is closed. Rather than maintain a constant open connection "state," each file that makes up the page (such as each of the graphics on a page) creates a new and separate connection (Privacy Foundation, 2001). This is why, for example, it is

sometimes possible to receive all the text of a Web page, but not the images; if the Web browser breaks the connection, or the distant server is too busy, it will not be able to open the additional connections needed to receive the additional data.

The benefit of a stateless connection is simple: It enables one machine to serve a much higher volume of data. The downside to a stateless connection is that on occasion it might be helpful for a server to remember who you are. For example, when someone logs onto his or her stock portfolio, privacy and security dictate that the server not reveal account information to anyone else; however, efficiency demands that every time the user loads a page, he or she should not have to reenter the user ID and password for every new connection the browser makes to the remote computer. So how do users make a server remember who they are? They do so by creating a constant state in an otherwise stateless series of connections. The method for doing this is the cookie.

Cookies contain a piece of data that allows the remote Web server to recognize a unique connection as having a relationship to another unique connection. In short, the cookie makes sure that the server can remember a visitor through many steps in a visit or even when time has passed between visits. As a basic security measure, it should be noted that cookies are designed to be read only by a server within the same domain that created it. So, for example, only a server in the yahoo.com domain can read cookies set by a server in the yahoo.com domain.

Cookies enable myriad helpful features, such as the ability to personalize a Web site with the user's choice of colors, or language, or stock symbols on a stock ticker. It also enables features such as shopping carts on e-commerce Web sites, permitting the user to select multiple items over the course of a long visit and have them queued for purchase at the end of a visit.

Not all cookies are used for collecting or retaining information over a long period of time, such as those used by advertisers. For example, many Web sites contain a great deal of frequently changing content and generate their Web pages from large databases of text. In some of these cases, the Web servers require cookies to help determine, for example, what page it should serve up to a user based on the search terms that he or she entered into a search engine.

A special type of cookie, called a session cookie, is set to be automatically deleted after a relatively short period of time, usually within about 10 minutes after a user leaves a site. This type of cookie is typically used for remembering information over a short duration, such as what you may have stored in a shopping cart. Because session cookies are so short-lived, they do not have quite the same privacy implications as their longer-lived cousin, the persistent cookie. Persistent cookies often have expiration dates set many years in the future.

Most Web browsers have settings that allow a user to accept or reject certain cookies. For example, an alternative brand of Web browser called Opera, favored among the privacy community, allows users to accept or reject cookies based on whether it is a first-party cookie being set by the site the user is actively visiting or whether it is a third-party cookie, which is being set by some other entity such as an advertising service via an ad banner appearing on the site.

Web Bugs

Another popular technology for tracking users' activities online is the Web bug, also called "Web beacons," "1-by-1 pixels," or "clear GIFs." (GIF, which stands for graphics interchange format, is a particular type of file format for images.)

Web bugs are special links imbedded in Web pages, or other HTML-coded documents such as some types of e-mail, that allow the link's creator to track every instance in which the document is viewed (Smith, 2001). As discussed earlier, every time a Web page is loaded, images on the page are loaded in a separate transaction with the Web server. When a Web bug is programmed into a Web page, its code looks similar to the code for just about any graphic image appearing on that page. In reality, though, it has three differences:

1. The Web bug graphic can be called from any site, most often from a third-party site, allowing that site to record details about the user's visit.
2. The Web address used to call in the Web bug graphic is often encoded with specific data relating to the page being visited, or, in the case of HTML e-mail, it may be encoded with information about the user's e-mail address.
3. The graphic image associated with the Web bug is deliberately made to be so tiny that it is invisible to the naked eye.

Most Web bugs are the size of a single screen pixel. What is a pixel? Every image on a computer screen is composed of very tiny dots. The smallest unit of dot on a computer screen is the pixel. Even a single pixel can still be visible, however, so Web bug images are often made of a graphic image called a clear GIF, or a transparent GIF, which allows the background color or image to show through it, rendering it effectively invisible.

Because Web bugs can be embedded in any Web page or HTML document, they can also be included in e-mail, allowing sites to track details about when a message is read and to whom the message might be sent. This versatility is why Web bugs have become so widely used. It is also why an industry group called the Network Advertising Initiative, which represents a growing category of online advertising firm called ad networks, responded to pressure from privacy advocates and legislators by agreeing to a set of guidelines for notice and choice when Web bugs are in use.

Ad Networks

Some sites rent out space on their Web pages to third parties, often for placement of advertisements. Along with those ad banners, many third-party advertising companies also try to set their own cookie on users' browsers. These cookies can be used for things such as managing ad frequency (the number of times an advertisement is shown to a particular individual) and to track users'

movements between the many sites on which the advertising companies place their cookies. These ad networks are a type of advertising agency that rents space on dozens or hundreds of Web sites, and frequently uses cookies placed on all of the sites in their network to build a profile about the kinds of Web sites a particular user likes to visit.

What is increasingly a marketer's paradise is becoming a consumer's nightmare: the deluge of commercial messages in e-mail inboxes, parades of pop-up advertisements, and even solicitations arriving by cellular phone and pager are making consumers leery of the alleged benefits of this ubiquitously wired world. In response to growing consumer concerns, companies have sought to develop privacy polices that help consumers better understand how their information is gathered and used.

PRIVACY POLICY FUNDAMENTALS

According to the Federal Trade Commission (FTC), if a company makes a promise that it does not keep, it is considered an unfair or deceptive trade practice, for which the offender can be fined up to $11,000 per violation, in addition to other legal remedies (FTC Office of the General Counsel, 2002). Central to the FTC's advocacy of greater consumer privacy protections has been the call for companies to adopt privacy policies that provide consumers with useful information about how their personal information is gathered and used. Although there are no federal laws that require the publication of a privacy policy, except when data collection from children is involved, it is widely considered an industry "best practice" to publish a privacy policy on any public Web site.

Considering the liability created by writing a privacy policy that a company cannot deliver, the drafting of a privacy policy is not something to be undertaken lightly or without advice of legal counsel. However, good privacy policies tend, at minimum, to address those elements contained in the widely accepted fair information principles, which have also been endorsed by the FTC: notice, choice, access, and security. I discuss the FTC's role in policing privacy matters later in this section, but it should also be noted that legal actions by state attorneys general, as well as private lawsuits, are also driving companies towards some level of uniformity in privacy disclosures.

Those privacy policies cited by privacy advocates as being "best of class" also include the elements of the OECD's principles of fair information practices. There are also a number of online privacy policy generators that allow one to create policies by picking and choosing from predefined language based on the applicable situation. According to the privacy organization, TRUSTe, their recommended Model Privacy Statement has several key elements that echo the OECD principles:

- What personally identifiable information the company collects,
- What personally identifiable information third parties collect through the Web site,
- What organization collects the information,
- How the company uses the information,
- With whom the company may share user information,
- What choices are available to users regarding collection, use, and distribution of the information,
- What types of security procedures are in place to protect the loss, misuse, or alteration of information under the company's control, and
- How users can correct any inaccuracies in the information (TRUSTe, 2002).

Once a company has surveyed its data practices and articulated them clearly in a privacy policy document, the next most important task is to ensure that the company lives up to its promises. There are three ways to do this: manage privacy matters internally, look to industry-sponsored groups for guidance on compliance, or wait for law enforcement to come after you.

Chief Privacy Officers

As the importance of privacy has grown in the corporate setting, and as the risks from privacy problems have increased, companies have begun to create a new management position, the chief privacy officer (CPO), as the designated point-person for managing privacy policies and practices.

Since the first CPO position was created in 1999 at the start-up Internet advertising firm AllAdvantage.com, the CPO job description (if not always the title) has been rapidly adopted across corporate America; by the end of 2000, a significant number of Fortune 100 firms had a CPO-type position, often reporting to the senior-most levels of the organization. According to the Privacy Working Group of the advocacy group Computer Professionals for Social Responsibility, there are many benefits to appointing a CPO:

> A talented and properly-positioned CPO will add value across corporate divisions from development to customer relations, from liability mitigation and risk management to increased market share and valuation. Perhaps most importantly, the Chief Privacy Officer promotes an essential element of new economy corporate citizenship-Trust. (Enright & McCullough, 2000)

The CPO has both an internal and an external role at his or her company. The internal role includes participation in companywide strategy planning, operations, product development and implementation, compliance monitoring and auditing, and employee training and awareness. The external role of the CPO involves enhancing the company's image as a privacy-sensitive organization, through fostering positive relationships with consumers and consumer groups, privacy advocates, industry peers, and regulators.

In many respect, the CPO becomes the focal point for a company's privacy activities and in turn can become the company's public face on the privacy issue. The position is most effective if it is perceived as objective, with ombudsman-like qualities, serving as a protector of consumer interests while seeking balance between those interests and the interests of the company. Yet there are other organizations offering assistance in the ombudsman role: trustmark organizations.

Trustmarks

There are several independent, industry-sponsored organizations that will certify a company's privacy policy to improve consumer perceptions. Upon certification, they permit sites to use their "seal of approval," sometimes referred to as a trustmark, to demonstrate to the public their commitment to privacy concerns. The most popular privacy seal programs—TRUSTe, BBBOnline, and CPA WebTrust—certify the validity of the policies on many thousands of Web sites. The growing use of these trustmarks does seem to be having an effect: an August 1999 study found that 69 percent of Internet users said that they recognize the TRUSTe seal, the most widely adopted of the privacy seal programs.

Seal programs verify that a Web site's privacy policy covers certain privacy topics (like the use of cookies and sharing data with third-party marketers). The seals do not set any specific quality standards, benchmarks, or specific data handling practices, however. As such, a company's site could, theoretically, earn a seal for making the required disclosures, even if in the course of the disclosure it reserves for itself the right to make whatever use of personal information it sees fit. This has been one of the criticisms leveled at the seal programs, as has their dependence on licensing fees from those entities they are asked to police.

Federal Trade Commission

Under their broad legislative mandate to proscribe deceptive and unfair trade practices, the FTC began reviewing online marketing practices back in 1996. Soon the FTC's investigators were uncovering evidence of some egregious behavior by a few online marketers. Major corporations quickly distanced themselves from the bad actors but acknowledged that privacy was a growing concern for online consumers and promised the FTC that the industry would do better at policing itself.

After numerous public controversies over well-known corporations continuing to abuse consumer privacy, and despite repeated pledges to adhere to standards promulgated and policed by the industry itself, surveys have continued to show that consumer perceptions of the potential for privacy abuses by online marketers continues to be a factor in consumer hesitance to embrace Internet commerce fully. In response, the FTC has sought on numerous occasions to assist companies in adopting practices that are more conducive to consumer confidence. These efforts have focused on the well-worn mantra: notice, choice, access, and security.

In December 1999, the FTC convened an Advisory Committee on Online Access and Security (ACOAS) to provide advice and recommendations to the agency regarding implementation of these basic fair information practices. The committee, consisting of representatives from the online industry, trade groups, academia, and privacy advocates, sought to provide guidance on how to solve the last two elements of fair information practices: access and security. Their report outlines many of the problems with setting universal standards for access and security, and in the end the committee came to few conclusions (FTC ACOAS, 2000).

Some six years after first looking into online privacy issues, the FTC is still warning online companies that if they do not clean up their act, stricter measures might be required. During the intervening years, however, the FTC has not been completely idle. In 1998, the FTC reached a settlement with GeoCities, a personal Web site hosting service, over charges that it misrepresented how user information would be used and engaged in deceptive practices relating to its collection of information from children. Part of the settlement required GeoCities (now owned by Yahoo!) to post, "a clear and prominent Privacy Notice, telling consumers what information is being collected and for what purpose, to whom it will be disclosed, and how consumers can access and remove the information." The notice, or a link to it, was required on the Web site's home page and on every page where information was collected (FTC, 1998).

In 1999, the FTC issued regulations implementing the Children's Online Privacy Protection Act, which requires online businesses to seek permission from parents before gathering personally identifiable information from children under age 13. It has since brought several enforcement actions to punish Web sites that have ignored those regulations (FTC, 2003). In recent years, the FTC has filed numerous actions against Internet-based fraudulent schemes, get-rich-quick scams, and quack medical remedies promoted via e-mail and on the Web. In 2000, the FTC intervened in the bankruptcy sale of a customer list belonging to defunct online toy retailer Toysmart.com. The basis of the action was to prevent the list from being used in any way inconsistent with the privacy policy under which it was gathered (FTC, 2000). In 2001, the FTC settled with pharmaceutical firm Eli Lilly and Company over an e-mail that improperly disclosed the e-mail addresses of hundreds of users of a prescription reminder service at the Web site Prozac.com, in violation of the site's privacy policy (FTC, 2002).

The FTC has steadfastly refused to seek greater legislative authority than its already broad mandate under the Federal Trade Act to police unfair or deceptive trade practices. The agency has threatened the industry that it will indeed seek more specific privacy-oriented enforcement authority if companies do not improve their self-regulatory efforts. It must also be noted that the FTC is just one governmental authority with the ability to prosecute privacy violations: Most state attorneys general have state versions of the Federal Trade Act that enable them to seek remedies similar to those available to the FTC. Indeed, attorneys general in Michigan, Washington, California, and Massachusetts have all been active in undertaking privacy-related enforcement actions, and the National Association of Attorneys General (2001) has held many seminars on investigating and prosecuting Internet privacy matters.

CONCLUSION

Consumers and businesses alike are grappling with the complex privacy concerns that the Internet era has brought to the fore. This chapter is a necessarily brief overview of the privacy landscape. Indeed, entire books can—and have–been written about the ways Internet

technologies have created new challenges to the average person's desire to "be let alone." As this chapter has shown, however, a number of concepts find their way into privacy-related policies and practices. Among these, the fundamental principles of notice, choice, access, and security are driving both consumer expectations and business planning. Keeping these principles in mind, many who are called on to seek privacy solutions in their own particular business or personal context have a conceptual framework within which to arrive at their own conclusion.

GLOSSARY

Ad network A consortium of Web sites linked together by an advertising agency for purposes of aggregating advertising placements and tracking consumers movements among and between member sites.

Cookies A small file saved by a Web browser, at the direction of a Web site, containing data that may be later retrieved by that Web site. See also persistent cookies, session cookies, third-party cookies.

Children's Online Privacy Protection Act of 1998 (COPPA) Legislation that limits operators of commercial Web sites and online services from collecting personal information from children under age 13.

Electronic Communications Privacy Act of 1986 (ECPA) Legislation governing the use of wiretaps for domestic law enforcement activities.

Foreign Intelligence Surveillance Act of 1978 (FISA) Legislation governing the use of wiretapping and physical searches in investigations involving terrorists and agents of foreign powers.

Gramm–Leach–Bliley Act (GLB) Also known as the Financial Services Modernization Act. Legislation that instituted major changes to the U.S. banking system. In pertinent part, GLB requires that organizations providing financial services disclose their data collection practices to customers and to provide the ability to opt-out of those practices.

Health Insurance Portability and Accountability Act of 1996 (HIPAA) Legislation that instituted a number of changes to health insurance practices. In pertinent part, HIPAA included privacy-related provisions applicable to health information created or maintained by health care providers, health plans, and health care clearinghouses.

Internet protocol (IP) address The unique numerical address assigned to each computer connected to the Internet. An address may be assigned temporarily (called a dynamic IP address) or may be assigned for long periods (called a static IP address).

Organization for Economic Cooperation and Development (OECD) A group of 30 democratic, market economy countries working collaboratively on economic, social, and trade issues.

Persistent cookies Cookie files designated to be stored for long periods, sometimes as long as 10 years.

Privacy Freedom from unauthorized intrusion. A state of being let alone and able to keep certain especially personal matters to oneself.

Safe harbor A legal concept that permits an entity to reduce or avoid legal liability by agreeing to adhere to certain standards or procedures. In the context of Internet privacy, safe harbor refers to an agreement between the United States and the European Union which permits U.S. companies to certify that they adhere to the stricter privacy standards required by European law, thereby avoiding a more burdensome set of country-by-country registration procedures.

Session cookies Cookie files designated to be stored for only the duration of a visit to a Web site; usually 10 minutes or less.

Third-party cookies A cookie file set by some entity other than the operator of the Web site being visited by the user. Third-party cookies are often used by advertising services to track user movements between multiple Web sites over periods of time.

Trustmark A symbol used to identify those companies whose Web sites have subjected their privacy policy to review by a third-party watchdog organization.

Web bugs Also called Web beacons, 1-by-1 pixels, or clear GIFs. Special links imbedded in Web pages, or other HTML-coded documents such as some types of e-mail, that allow the link's creator to track every instance in which the document is viewed.

CROSS REFERENCES

See *Cyberlaw: The Major Areas, Development, and Provisions; International Cyberlaw; Legal, Social and Ethical Issues.*

REFERENCES

American Management Association (2001). 2001 workplace monitoring and surveillance: Policies and practices. Retrieved May 9, 2003, from http://www.amanet.org/research/archives.htm

Cailliau, R. (1995). *A little history of the World Wide Web*. Retrieved December 3, 2002, from http://www.w3c.org/History.html

Cantrell v. Forest City Publishing Co., 419 U.S. 245 (1974). Retrieved December 3, 2002, from http://laws.findlaw.com/us/419/245.html

Children's Online Privacy Protection Act, 15 U.S.C. §§6501—6506 (1998). Retrieved December 3, 2002, from http://www.ftc.gov/ogc/coppa1.htm

Douglass v. Hustler Magazine, 769 F.2d 1128 (1985).

Electronic Communications Privacy Act, 18 U.S.C. §2701 (1986). Retrieved December 3, 2002, from http://www4.law.cornell.edu/uscode/18/2701.html

Electronic Privacy Information Center Carnivore Archive (2002). Retrieved February 9, 2003, from http://www.epic.org/privacy/carnivore/

Electronic Privacy Information Center Foreign Intelligence Surveillance Act Archive (2003). Retrieved February 8, 2003, from http://www.epic.org/privacy/terrorism/fisa/

Electronic Privacy Information Center Total Information Awareness Archive (2003). Retrieved February 8, 2003, from http://www.epic.org/privacy/profiling/tia/

Enright, K. P., & McCullough, M. R. (2000). *Computer professionals for social responsibility privacy working group: CPO guidelines*. Retrieved December 3, 2002, from http://www.privacylaw.net/CPO_Guidelines.pdf

European Commission (1995). Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data. *Official Journal L., 281,* 31. Retrieved December 3, 2002, from http://europa.eu.int/comm/internal_market/en/dataprot/law/

European Union Parliament (2001). Temporary Committee on the ECHELON Interception System Report. Retrieved February 9, 2003, from http://www.europarl.eu.int/tempcom/echelon/pdf/prechelon_en.pdf

Fair Credit Reporting Act (15 U.S.C. §§1681–1681(u), as amended). Retrieved December 3, 2002, from http://www.ftc.gov/os/statutes/fcra.htm

Federal Trade Commission (1998). In the matter of Geo Cities, File No. 982 3051. Retrieved December 3, 2002, from http://www.ftc.gov/opa/1998/9808/geocitie.htm

Federal Trade Commission (2000). FTC v. Toysmart.com, LLC, and Toysmart.com, Inc. (District of Massachusetts) (Civil Action No. 00-11341-RGS). Retrieved December 3, 2002, from http://www.ftc.gov/opa/2000/07/toysmart.htm

Federal Trade Commission (2002). In the Matter of Eli Lilly and Company, File No. 012 3214. Retrieved December 3, 2002, from http://www.ftc.gov/opa/2002/01/elililly.htm

Federal Trade Commission Advisory Committee on Online Access and Security (2000). *Final report*. Retrieved December 3, 2002, from http://www.ftc.gov/acoas

Federal Trade Commission (2003). "Kidz Privacy" education campaign site. Retrieved February 2, 2003, from http://www.ftc.gov/bcp/conline/edcams/kidzprivacy/news.htm

Federal Trade Commission Office of the General Counsel (2002). A brief overview of the Federal Trade Commission's investigative and law enforcement authority. Retrieved February 2, 2003, from http://www.ftc.gov/ogc/brfovrvw.htm

Foreign Intelligence Surveillance Act of 1978 (codified at 50 U.S.C. §§1801–1811, 1821–1829, 1841–1846, 1861–62, as amended). Retrieved December 3, 2002, from http://www4.law.cornell.edu/uscode/50/1801.html

Freedom of Information Act of 1966 (5 U.S.C. §552). Retrieved December 3, 2002, from http://www.usdoj.gov/oip/foia_updates/Vol_XVII_4/page2.htm

Gramm–Leach–Bliley Act (codified in relevant part at 15 U.S.C. §§6801–6809). Retrieved December 3, 2002, from http://www.ftc.gov/privacy/glbact/glbsub1.htm

Levine, J. R., Everett-Church, R., & Stebben, G. (2002). Internet Privacy for Dummies. New York: Wiley Publishing, Inc.

McCullagh, D. (2003). Labels, Verizon DMCA battle rages. *ZDNet News*. Retrieved February 12, 2003, from http://zdnet.com.com/2100-1104-983896.html

National Association of Attorneys General (2001). 39 attorneys general, the District of Columbia corporation counsel, and the Georgia Governors Office of Consumer Affairs submit comments to FCC urging better privacy protections for consumers. Retrieved December 3, 2002, from http://www.naag.org/issues/pdf/20011228-signon-fcc.pdf

Network Advertising Initiative (NAI) (2001). Web Beacons: Guidelines for Notice and Choice. Retrieved December 3, 2002, from http://www.networkadvertising.org/Statement.pdf

Oakes, C. (1999). 23 fired for e-mail violations. *Wired News*. Retrieved December 3, 2002, from http://www.wired.com/news/politics/0,1283,32820,00.html

Olmstead v. United States, 277 U.S. 438 (1928). Retrieved December 3, 2002, from http://laws.findlaw.com/us/277/438.html

Organization for Economic Cooperation and Development (1980). *Guidelines on the protection of privacy and transborder flows of personal data*. Retrieved December 3, 2002, from http://www.oecd.org/EN/document/0,,EN-document-43-1-no-24-10255-43,00.html

Privacy Act of 1974 (5 U.S.C. §552A). Retrieved December 3, 2002, from http://www.usdoj.gov/foia/privstat.htm

Privacy Foundation (2001). Cookie Maker. Retrieved December 3, 2002, from http://www.privacyfoundation.org/resources/montulli.asp

Prosser, W. L. (1960). Privacy. *California Law Review, 48,* 383.

Smith, R. M. (2001). FAQ: Web bugs. Retrieved December 3, 2002, from http://www.privacyfoundation.org/resources/webbug.asp

Sullivan, B. (2001). FBI software cracks encryption wall. MSNBC News Web site. Retrieved February 8, 2003, from http://www.msnbc.com/news/660096.asp

TRUSTe (2002). Model privacy statement. Retrieved December 3, 2002, from http://www.truste.org/webpublishers/pub_modelprivacystatement.html

United Nations (1948). Universal Declaration of Human Rights. Retrieved December 3, 2002, from http://www.un.org/Overview/rights.html

U.S. Department of Commerce (2000). *Safe harbor*. Retrieved December 3, 2002, from http://www.export.gov/safeharbor

U.S. Department of Health, Education and Welfare (1973). Records, computers and the rights of citizens. Retrieved December 3, 2002, from http://aspe.hhs.gov/datacncl/1973privacy/tocprefacemembers.htm

U.S. Department of Health and Human Services (2002). *Medical privacy—national standards to protect the privacy of personal health information, Office for Civil Rights*. Retrieved December 3, 2002, from http://www.hhs.gov/ocr/hipaa/

Uniting and Strengthening America by Providing Appropriate Tools Required to Intercept and Obstruct Terrorism (USA Patriot) Act of 2001 (Pub. L. No. 107–56). Retrieved December 3, 2002, from http://www.epic.org/privacy/terrorism/hr3162.html

Project Management Techniques

Kathy Schwalbe, *Augsburg College*

INTRODUCTION

Although projects have been managed in some fashion for thousands of years, changes in society, the workforce, and technology have sparked interest in the topic of modern project management. The Project Management Institute (PMI), a professional society with more than 100,000 members worldwide, estimates that the United States spends more than $2.3 trillion on projects every year, or one quarter of the nation's gross domestic product (PMI, 2001a). Many of these projects involved information technology and the Internet. According to the Standish Group, there has been an information technology "project gold rush." In 1998, corporate America issued 200,000 new-start application software development projects. During 2000, there were 300,000 new starts, and more than half a million new start application software development projects were initiated during 2001 (Standish Group, 2001). These changes have fueled the need for more sophisticated and better project management. In fact, today's corporations are recognizing that to be successful, they need to be conversant with and use modern project management techniques.

What Is a Project?

To discuss project management techniques, it is important to first understand the concept of a project. A project is a temporary endeavor undertaken to accomplish a unique purpose. Projects should be aligned with organizational objectives. For example, if an organization is trying to develop new products, decrease time to market, increase revenues, or cut costs, projects should be initiated to support those goals. The following attributes help to further define a project:

A project has a unique purpose. Every project should have a well-defined objective. For example, a company might want to develop an intranet to streamline internal operations and cut costs. They might estimate that they can save thousands of dollars on printing and training costs by putting information on a well-designed Intranet. The project purpose must state clearly what outcomes and deliverables are expected from the project, which departments and processes in the organization are involved, and so on.

A project is temporary. A project has a definite beginning and a definite end. If a company is doing a project to develop an intranet, it should decide when the project is expected to begin and end. The project team should define and then produce deliverables along the way, such as a concept document, preliminary design, test site, and so on. The temporary nature of projects most distinguishes them from normal operations in an organization.

A project requires resources, often from various areas. Resources include people, hardware, software, or other assets. Many projects cross departmental or other boundaries to achieve their unique purposes. For an intranet project, people from information technology, marketing, sales, distribution, and other areas of a company would need to work together to develop ideas. They might also hire outside consultants to provide input on the design of the intranet and which technologies to use for the project. Once the project team finishes its intranet design, they might find that the company requires additional hardware, software, and network resources. People from other companies—product vendors and consulting companies—would also become resources for meeting project objectives. Resources, however, are not unlimited. They must be used effectively in order to meet project and other corporate goals.

A project should have a primary sponsor or customer. Most projects have many interested parties or stakeholders, but someone must take the primary role of sponsorship. The project sponsor usually provides the direction

and funding for the project and helps make major decisions.

A project involves uncertainty. Because every project is unique, it is sometimes difficult to clearly define the project's objectives, estimate how long it will take to complete, or how much it will cost. This uncertainty is one of the main reasons project management is so challenging, especially on projects involving new technologies.

Every project is also constrained in different ways by its scope, time goals, and cost goals. These limitations are sometimes referred to in project management as the triple constraint. To create a successful project, scope, time, and cost must all be taken into consideration, and it is the project manager's duty to balance these three often-competing goals. Project managers must consider the following:

Scope: What is the project trying to accomplish? What unique product or service does the customer or sponsor expect from the project? What is and is not included in the project scope? Who will verify the project scope and each deliverable?

Time: How long should it take to complete the project? What is the project's schedule? Which tasks have dependencies on other tasks? What is driving the project completion date?

Cost: What should it cost to complete the project? How will costs be tracked? How will cost variances be handled?

Project success is often measured by meeting scope, time, and cost goals. Note that each area—scope, time, and cost—has a target at the beginning of the project. For example, an intranet project might initially involve six departments, an estimate of one hundred Web pages, and four internal database applications. The project team might estimate that developing this new intranet will save the organization $300,000 within 1 year after implementation. The initial time estimate for this project might be 9 months, and the cost estimate might be $200,000. These expectations would provide the targets for the scope, time, and cost dimensions of the project as well as the organizational benefits.

Project teams do want to meet scope, time, and cost goals, but they must also focus on satisfying project stakeholders and supporting organizational objectives. Project management, therefore, involves several other dimensions or knowledge areas.

What Is Project Management?

Project management is "the application of knowledge, skills, tools, and techniques to project activities in order to meet project requirements" (PMBOK® Guide, 2000 Edition, 2000, p. 4). Project managers must not only strive to meet specific scope, time, and cost goals of projects, they must also meet quality goals and facilitate the entire process to meet the needs and expectations of the people involved in or affected by project activities.

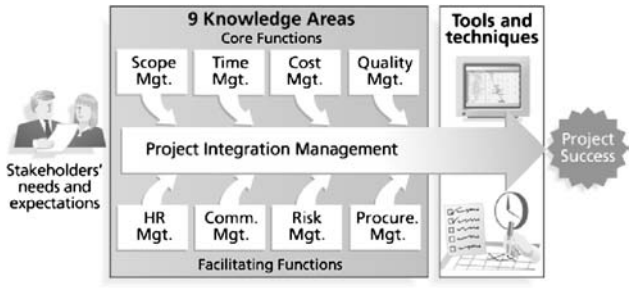


Figure 1: Project management framework.

Figure 1 provides a framework for beginning to understand project management. Key elements of this framework include the project stakeholders, project management knowledge areas, and project management tools and techniques.

Stakeholders are the people involved in or affected by project activities and include the project sponsor, project team, support staff, customers, users, suppliers, and even opponents to the project. People's needs and expectations are important in the beginning and throughout the life of a project. Successful project managers work on developing good relationships with project stakeholders to ensure their needs and expectations are understood and met.

Knowledge areas describe the key competencies that project managers must develop. The center of Figure 1 shows the nine knowledge areas of project management. The four core knowledge areas of project management include project scope, time, cost, and quality management. These are considered to be core knowledge areas because they lead to specific project objectives. Brief descriptions of each follow.

1. *Project scope management* involves defining and managing all the work required to successfully complete the project.
2. *Project time management* includes estimating how long it will take to complete the work, developing an acceptable project schedule and ensuring timely completion of the project.
3. *Project cost management* consists of preparing and managing the budget for the project.
4. *Project quality management* ensures that the project will satisfy the stated or implied needs for which it was undertaken.

The four facilitating knowledge areas of project management are human resources, communications, risk, and procurement management. These are called facilitating areas because they are the means through which the project objectives are achieved. Brief descriptions of each follow.

1. *Project human resource management* is concerned with making effective use of the people involved with the project.
2. *Project communications management* involves generating, collecting, disseminating, and storing project information.

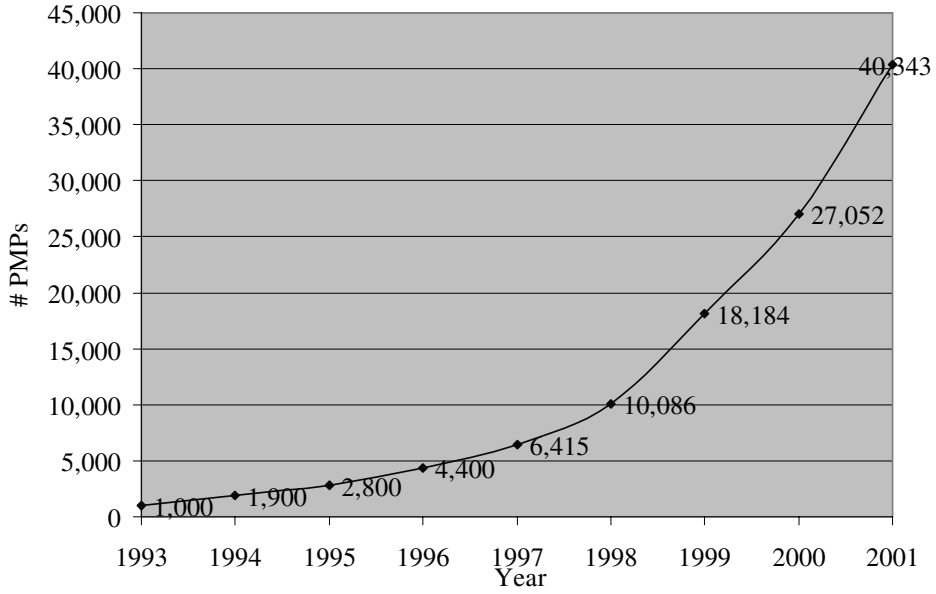


Figure 2: Growth in Project Management Professional certification, 1993–2001.

3. *Project risk management includes* identifying, analyzing, and responding to risks related to the project.
4. *Project procurement management* involves acquiring or procuring goods and services that are needed for a project from outside the performing organization.

Project integration management, the ninth knowledge area, is an overarching function that affects and is affected by all of the other knowledge areas. Project managers must have knowledge and skills in all nine of these areas.

BRIEF BACKGROUND ON THE PROJECT MANAGEMENT PROFESSION

Most people in the information technology field work on projects in some capacity. Some people become project managers, at least on a part-time basis, early in their careers. Many people think one needs many years of experience and vast technical knowledge to be a project manager. Although this need for experience and knowledge might be the case for large, complex, and expensive projects, many information technology projects can be, and are, led by project managers who are just starting their careers. Project managers need some general management and technical knowledge, but they primarily need the skills and desire to be project managers. Project team members should also understand project management to contribute effectively to projects.

The project management profession is growing at a rapid pace. Fortune magazine called project management the "number one career choice" in its article "Planning a Career in a World Without Managers" (Stewart & McGowan, 1996). PMI, an international professional society for project managers, estimated that the average salary for a project manager was more than $81,000 (PMI, 2000a). Although many professional societies are suffering from declining memberships, PMI membership continues to grow at a rapid pace, and the organization included more than 100,000 members worldwide in early 2003. A large percentage of PMI members work in the information technology field, and many pay additional dues to join the Information Systems Specific Interest Group.

Professional certification is an important factor in recognizing and ensuring quality in a profession. PMI provides certification as a Project Management Professional (PMP)—someone who has documented sufficient project experience and education, agreed to follow the PMI code of ethics, and demonstrated knowledge of the field of project management by passing a comprehensive examination. The number of people earning PMP certification continues to increase. In 1993, there were about 1,000 certified project management professionals. By the end of 2001, there were more than 43,000 certified project management professionals. Figure 2 shows the rapid growth in the number of people earning project management professional certification from 1993 to 2001, based on data provided in PMI's annual reports.

As information technology projects become more complex and global in nature, the need for people with demonstrated knowledge and skills in project management will continue. Just as passing the certified public accountant exam is a standard for accountants, passing the PMP exam is becoming a standard for project managers. Some companies are requiring that all project managers be PMP certified. Project management certification is also enabling professionals in the field to share a common base of knowledge. For example, any PMP should be able to describe the nine project management knowledge areas and apply many of the tools and techniques discussed in this chapter. Sharing a common base of knowledge is important because it helps advance the theory and practice of project management.

KEY PROJECT MANAGEMENT TOOLS AND TECHNIQUES

There are many tools and techniques available to project managers and their teams to assist them in all of the knowledge areas. This section highlights just a few of them. Consult the PMBOK Guide (PMI, 2000b), *Information Technology Project Management* (Schwalbe, 2002), or other references for more detailed information.

Project Selection Techniques

Some of the most important decisions organizations must make involve which projects to pursue. It does not matter if a project is highly successful in meeting scope, time, and cost goals if the project itself is not important to the organization. Therefore, organizations should develop and follow a logical process for selecting and continuing projects. Four common techniques for selecting projects include the following:

1. *Focusing on broad organizational goals:* Organizations often cite broad needs, such as improving customer relationships, increasing market share, and so on. Senior managers often like to see strong financial justification for projects, but sometimes it is sufficient to show that projects support high-level strategies that meet broad organizational needs.
2. *Categorizing projects:* Not all projects can be high priority or tied to meeting a critical corporate goal. Projects are often started to address problems, opportunities, or directives that arise. Projects can also be categorized based on the resources they need in terms of time and cost.
3. *Performing net present value and other financial analyses:* Financial considerations are often an important aspect of the project selection process, especially in tough economic times. The primary methods for determining projected financial value of projects include net present value analysis, return on investment, and payback analysis. These techniques are not unique to project management. Any financial analyst can describe these techniques and their importance to senior managers. Consult a good finance or project management textbook or Web site (http://www.investopedia.com) for descriptions of these techniques.
4. *Using a weighted scoring model:* A weighted scoring model is a tool that provides a systematic process for selecting projects based on many criteria. These criteria can include factors such as meeting broad organizational needs; addressing problems, opportunities, or directives; the amount of time it will take to complete the project; the overall priority of the project; projected financial performance of the project; and so on.

The first step in creating a weighted scoring model is to identify criteria important to the project selection process. It often takes time to develop and reach agreement on these criteria. Holding facilitated brainstorming sessions or using groupware to exchange ideas can aid in developing these criteria. Some possible criteria for information technology projects include the following:

- Supports key business objectives
- Has strong internal sponsor
- Has strong customer support
- Uses realistic level of technology
- Can be implemented in 1 year or less
- Provides positive net present value
- Has low risk in meeting scope, time, and cost goals

Next, the manager assigns a weight to each criterion. These weights indicate how much stakeholders value each criterion or how important each criterion is. Weights can be assigned based on percentages, and the total of all of the criteria's weights must total 100%. The manager then assigns numerical scores to each criterion (for example, 0 to 100) for each project. The scores indicate how much each project meets each criterion. At this point, the manager can use a spreadsheet application to create a matrix of projects, criteria, weights, and scores. Figure 3 provides an example of a weighted scoring model to evaluate four projects. After assigning weights for the criteria and scores for each project, the manager calculates a weighted score for each project by multiplying the weight for each criterion by its score and adding the resulting values.

For example, one could calculate the weighted score for Project 1 in Figure 3 as follows:

$$25\% * 90 + 15\% * 70 + 15\% * 50 + 10\% * 25 + 5\% * 20 + 20\% * 50 + 10\% * 20 = 56.$$

Note that in this example, Project 2 would be the obvious choice for selection because it has the highest weighted score. Creating a bar chart to graph the weighted scores for each project allows one to see the results at a glance. If a manager creates the weighted scoring model in a spreadsheet, he or she can enter the data, create and copy formulas, and perform "what if" analysis. For example, suppose the manager changes the weights for the criteria. By having the weighted scoring model in a spreadsheet, he or she can easily change the weights, and the weighted scores and charts are updated automatically.

Many organizations are striving to select and manage projects better by using project portfolio management, enterprise project management, and multiproject management techniques, all topics beyond the scope of this chapter.

Formalizing Projects With a Project Charter

Once an organization has decided which projects to pursue, it is important to authorize those projects formally and inform other people in the organizations about the project objectives, schedule, and so on. One common tool used for this project authorization is a project charter. Project charters can take several forms, such as a simple letter of agreement, a formal contract, or an organization's project charter format. Key project stakeholders should sign the project charter to acknowledge agreement on the need and intent of the project. A project charter is a key output of project initiation.

Table 1 provides a sample of a project charter for upgrading a company's information technology

	A	B	C	D	E	F
1	**Criteria**	**Weight**	**Project 1**	**Project 2**	**Project 3**	**Project 4**
2	Supports key business objectives	25%	90	90	50	20
3	Has strong internal sponsor	15%	70	90	50	20
4	Has strong customer support	15%	50	90	50	20
5	Realistic level of technology	10%	25	90	50	70
6	Can be implemented in one year or less	5%	20	20	50	90
7	Provides positive NPV	20%	50	70	50	50
8	Has low risk in meeting scope, time, and cost goals	10%	20	50	50	90
9	**Weighted Project Scores**	**100%**	**56**	**78.5**	**50**	**41.5**
10						

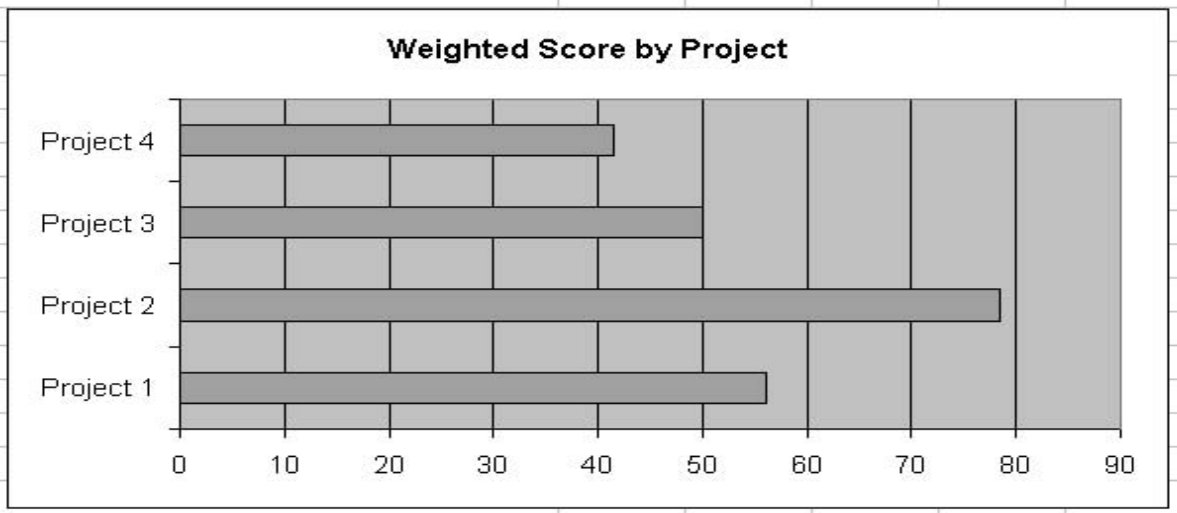


Figure 3: Sample weighted scoring model for project selection.

infrastructure. Because many projects fail because of unclear requirements and expectations, starting with a simple project charter makes a lot of sense.

Defining Project Scope With a Scope Statement and Work Breakdown Structure

A project charter helps define the high-level scope of a project, but much more work must be done to clarify the work to be done. Two important tools for further defining project scope are the scope statement and the work breakdown structure. A scope statement is a document used to develop and confirm a common understanding of the project scope. The scope statement should include a project justification, a brief description of the project's products, a summary of all project deliverables, and a statement of what determines project success. Scope statements vary significantly by type of project. Large, complex projects have long scope statements. Government projects or contracts often include a scope statement known as a statement of work (SOW). Some SOWs are hundreds of pages long, particularly if they include detailed product specifications. As with many other project management documents, the scope statement should be tailored to meet the needs of the particular project.

After completing scope planning, the next step in project scope management is to further define the work required for the project and to break it into manageable pieces. Breaking work into manageable pieces is called scope definition. Good scope definition is important to project success because it helps improve the accuracy of time, cost, and resource estimates, it defines a baseline for performance measurement and project control, and it aids in communicating clear work responsibilities. The output of the scope definition process is the work breakdown structure for the project.

A work breakdown structure (WBS) is a deliverable-oriented grouping of the work involved in a project that defines the total scope of the project. A WBS shows work decomposition, meaning each descending level shows a more detailed breakout of the level above it. It is a foundation document in project management because it provides the basis for planning and managing project schedules, costs, and changes. Project management experts believe that work should not be done on a project if it is not included in the WBS. PMI recently published its first practice standard on the topic of work breakdown structures (PMI, 2001) due to member requests for more guidance on developing and applying this important project management tool.

A WBS is often depicted as a task-oriented family tree of activities. It is usually organized around project products or by phases. It can also be organized using the project management process groups. It can look like an organizational chart, which can help people visualize the whole project and all of its main parts. A WBS can also

Table 1 Sample Project Charter

Project Title: Information Technology (IT) Upgrade Project

Project Start Date: March 4, 2002 **Projected Finish Date:** December 4, 2002

Project Manager: Kim Nguyen, 691–2784, knguyen@abc.com

Project Objectives: Upgrade hardware and software for all employees (approximately 2,000) within 9 months based on new corporate standards. See attached sheet describing the new standards. Upgrades may affect servers and midrange computers, as well as network hardware and software. Budgeted $1,000,000 for hardware and software costs and $500,000 for labor costs.

Approach:

- Update the information technology inventory database to determine upgrade needs
- Develop detailed cost estimate for project and report to CIO
- Issue a request for quotes to obtain hardware and software
- Use internal staff as much as possible to do the planning, analysis, and installation

Roles and Responsibilities:

NAME	ROLE	RESPONSIBILITY
Walter Schmidt, CEO	Project Sponsor	Monitor project
Mike Zwack	CIO	Monitor project, provide staff
Kim Nguyen	Project Manager	Plan and execute project
Jeff Johnson	Director IT Operations	Mentor Kim
Nancy Reynolds	VP, Human Resources	Provide staff, issue memo to all employees about project
Steve McCann	Director of Purchasing	Assist in purchasing hardware and software

Signoff: (Signatures of all the above stakeholders)

Comments: (Handwritten comments from above stakeholders, if applicable)

This project must be done within 10 months at the absolute latest. Mike Zwack, CIO.

We are assuming that adequate staff will be available and committed to supporting this project. Some work must be done after hours to avoid work disruptions, and overtime will be provided. Jeff Johnson and Kim Nguyen, Information Technology Department

Note. From Schwalbe (2002). © 2002 by Course Technology, a division of Thompson Learning. Reprinted with permission.

be shown in tabular form as an indented list of tasks to show the same groupings of the work. Figure 4 shows a WBS for an intranet project. Notice that product areas provide the basis for its organization. In this case, there is a main box or item on the WBS for developing the Web site design, the home page for the intranet, the marketing department's pages, and the sales department's pages.

In contrast, a WBS for the same intranet project can be organized around project phases, as shown in Figure 5. Notice that project phases of concept, Web site design, Web site development, roll out, and support provide the basis for its organization.

This same WBS is shown in tabular form in Table 2. The items on the WBS are the same, but the numbering scheme and indentation of tasks show the structure. This tabular form is used in many documents, such as contracts. It is also used in project management software, such as Microsoft Project.

The work breakdown structures in Figures 4, 5, and 6 and in Table 2 present information in hierarchical form. The top level of a WBS is the 0 level and represents the entire project. (Note the labels on the left side of Figure 5). The next level down is Level 1, which represents the major

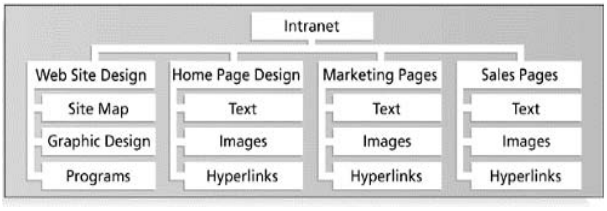


Figure 4: Sample intranet work breakdown structure organized by product.

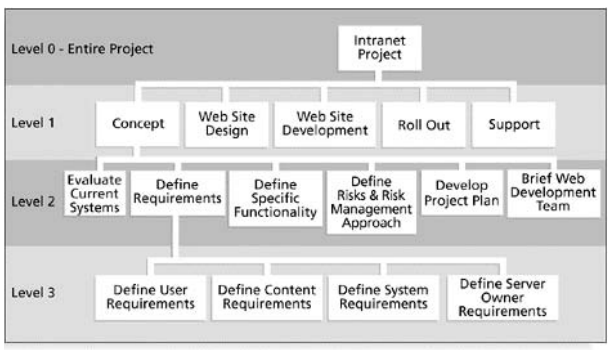


Figure 5: Sample intranet work breakdown structure organized by phase.

Table 2 Intranet Work Breakdown Structure in Tabular Form

1.0 Concept
1.1 Evaluate current systems
1.2 Define requirements
1.2.1 Define user requirements
1.2.2 Define content requirements
1.2.3 Define system requirements
1.2.4 Define server owner requirements
1.3 Define specific functionality
1.4 Define risks and risk management approach
1.5 Develop project plan
1.6 Brief web development team
2.0 Web site design
3.0 Web site development
4.0 Roll out
5.0 Support

Note: From Schwalbe (2002). © 2002 by Course Technology, a division of Thompson Learning. Reprinted with permission. Figure 6 shows this phase-oriented intranet WBS, using the same numbering scheme, in Microsoft Project 2000. One can see from this figure that the WBS is the basis for project schedules. Notice that the WBS is in the left part of the figure, and the resulting schedule in the form of a Gantt chart is in the right part of the figure.

products or phases of the project. Level 2 includes the major subsets of Level 1. For example, in Figure 5 the Level 2 items under the Level 1 item called "Concept" include the following: evaluate current system, define requirements, define specific functionality, define risks and risk management approach, develop project plan, and brief Web development team. Under the Level 2 item called "Define Requirements" are four Level 3 items on the WBS: define user requirements, define content requirements, define server requirements, and define server owner requirements.

In Figure 5, the lowest level provided is Level 3. The lowest level of the WBS represents work packages. A work package is a deliverable or product at the lowest level of the WBS. As a general rule, each work package in a WBS should represent roughly 80 hours of effort. One can also think of work packages in terms of accountability and reporting. The WBS should partition the work to allow semiautonomous work assignment for individuals and teams. If a project has a relatively short time frame and requires weekly progress reports, a work package might represent 40 hours of work. On the other hand, if a project has a long time frame and requires quarterly progress reports, a work package might represent more than 100 hours of work.

The sample WBSs shown here seem fairly easy to construct and understand. Nevertheless, it is difficult to create a good WBS. To create a good WBS, a project manager must understand both the project and its scope and incorporate the needs and knowledge of stakeholders. The manager and the project team must decide as a group how to organize the work and how many levels to include in the WBS. Many project managers have found that it is better to focus on getting the top three levels done well before getting too bogged down in more detail.

Another concern when creating a WBS is how to organize it so it provides the basis for the project schedule.

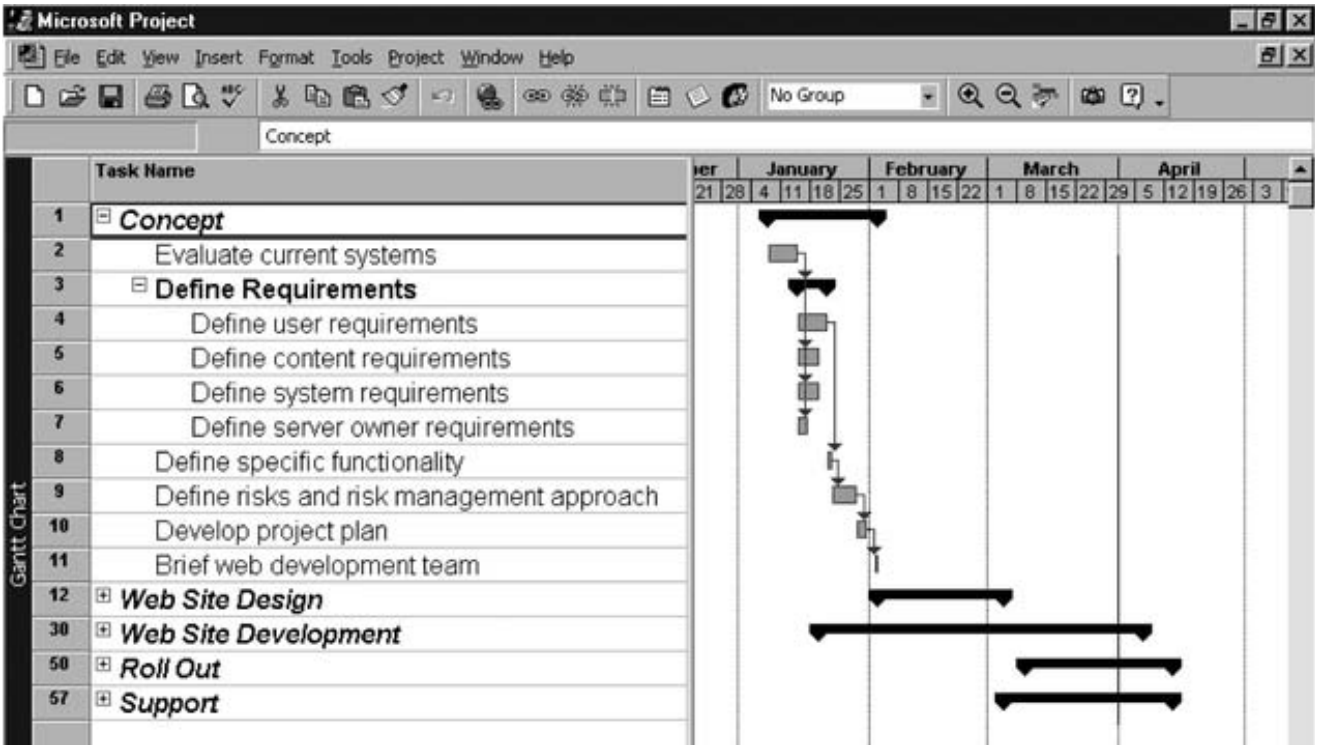


Figure 6: Intranet work breakdown structure and Gantt chart in Microsoft Project.

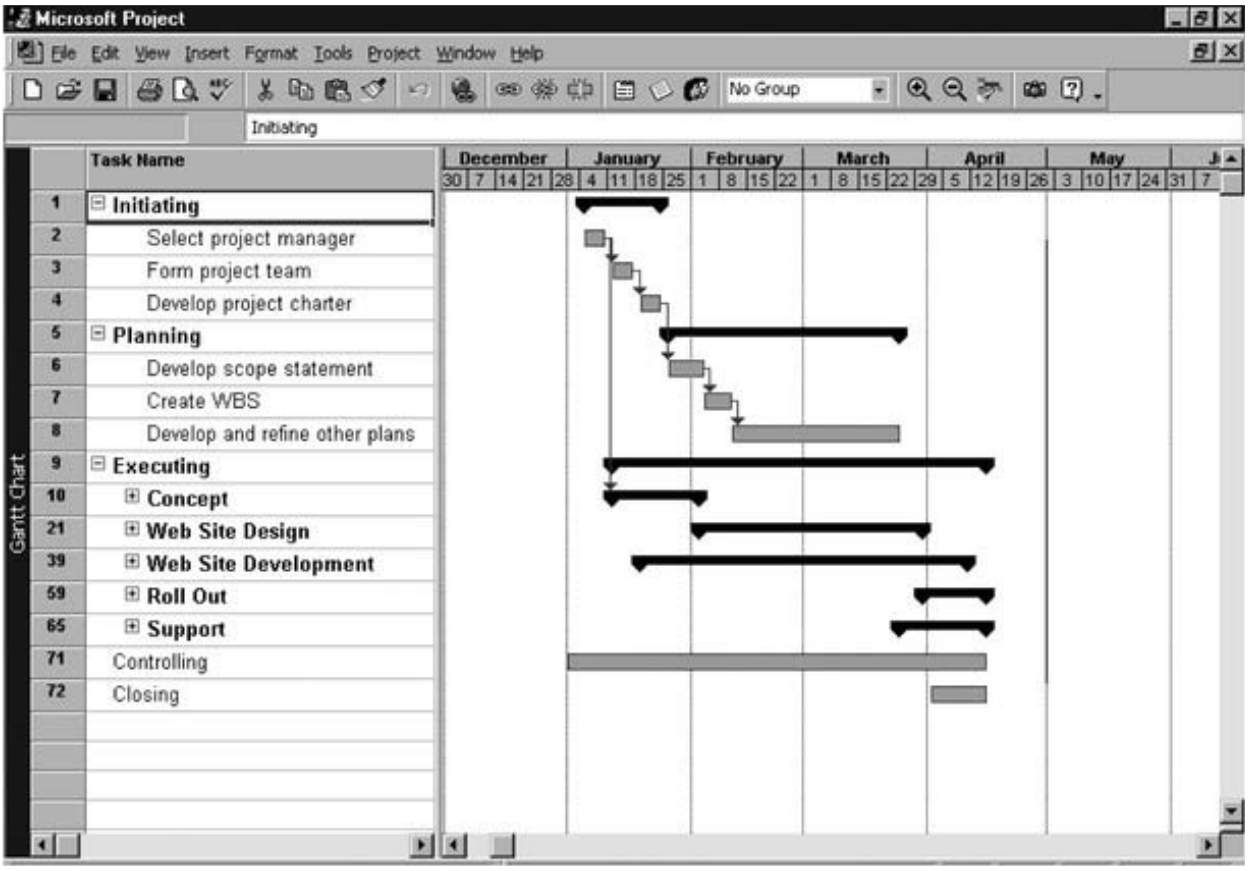


Figure 7: Intranet work breakdown structure and Gantt chart organized by project management process groups.

Some managers suggest creating a WBS using the project management process groups of initiating, planning, executing, controlling, and closing as Level 1 in the WBS. By doing this, not only does the project team follow good project management practice, but the WBS tasks can also be mapped more easily against time. For example, Figure 7 shows a WBS and Gantt chart for the intranet project organized by the product phases described earlier. Tasks under initiating include selecting a project manager, forming the project team, and developing the project charter. Tasks under planning include developing a scope statement, creating a WBS, and developing and refining other plans, which would be broken down in more detail for a real project. The tasks of concept, Web site design, Web site development, and roll out, which were WBS Level 1 items in Figure 6, now become WBS Level 2 items under executing. The executing tasks would vary the most from project to project, but many of the tasks under the other project management process groups would be similar for all projects.

It is important to involve the entire project team and customer in creating and reviewing the WBS. People who will do the work should help to plan the work by creating the WBS. Having group meetings to develop a WBS helps everyone understand what work must be done for the entire project and how it should be done, given the people involved. It also helps to identify where coordination between different work packages will be required.

Assigning Resources With a Responsibility Assignment Matrix

There are many tools and techniques to assist in managing people on projects. One simple yet effective tool to help clarify roles and responsibilities is the responsibility assignment matrix (RAM). A responsibility assignment matrix (RAM) is a matrix that maps the work of the project as described in the WBS to the people responsible for performing the work as described in an organizational breakdown structure (OBS). Figure 8 shows an example of a RAM. The positions of the WBS and OBS can be reversed, if desired. The RAM allocates work to responsible and performing organizations, teams, or individuals, depending on the desired level of detail. For smaller projects, it would be best to assign individual people to WBS activities. For large projects, it is more effective to assign the work to organizational units or teams.

In addition to using a RAM to assign detailed work activities, project managers can also use a RAM to define general roles and responsibilities on projects. This type of RAM can include the stakeholders in the project. Figure 9 provides a RAM that shows whether stakeholders are accountable or just participants in part of a project and whether they are required to provide input, review, or sign off on parts of a project. This simple tool can be an effective way for project managers to communicate roles and expectations of important stakeholders on projects.

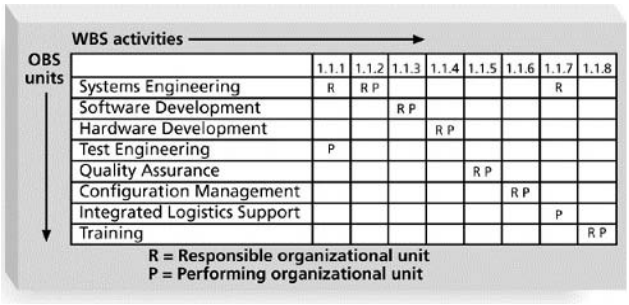


Figure 8: Sample responsibility assignment matrix.

PROJECT SCHEDULING TOOLS AND TECHNIQUES

When many people think of project management tools and techniques, they think of Gantt charts and network diagrams. Of course all nine knowledge areas are important in project management, but scheduling problems often cause the most conflicts on projects. This section briefly describes Gantt charts, network diagrams, and a more recent development in project scheduling called critical chain scheduling.

Gantt Charts

Figures 6 and 7 show how a WBS is a basis for creating a Gantt chart. Project managers must know what work needs to be done in order to develop a project schedule. Gantt charts provide a standard format for displaying project schedule information by listing project activities and their corresponding start and finish dates in a calendar format. Henry Gantt developed the first Gantt chart during World War I for scheduling work in job shops. Early versions simply listed project activities or tasks in one column to the left, calendar time units such as months to the right, and horizontal bars under the calendar units to illustrate when activities should start and end. Gantt charts normally do not show relationships between project activities as network diagrams do, however.

Today most people use project management software to create more sophisticated versions of Gantt charts and allow for easy updates of information. This chapter includes several figures created with Microsoft Project 2000. There are many other project management software tools available (see http://www.infogoal.com/pmc/pmcswr.htm or www.allpm.com). Note that project management software can do much more than just create Gantt charts.

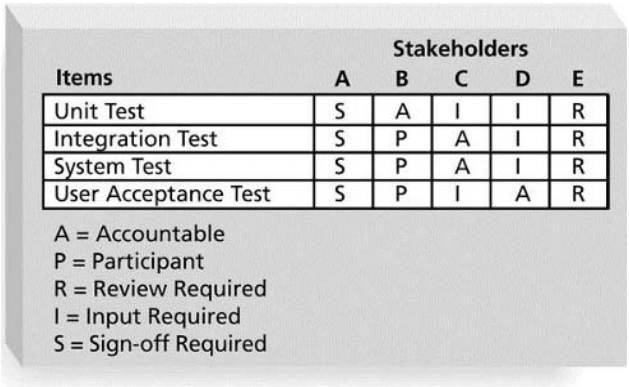


Figure 9: Responsibility assignment matrix showing stakeholder roles.

Figure 10 shows a Gantt chart based on a software launch project, one of the template files that Microsoft provides with Microsoft Project. Recall that the activities or items in the Task column on the left side of the figure coincide with the activities on the WBS for the project. Notice that the software launch project's Gantt chart contains milestones, summary tasks, individual task durations, and arrows showing task dependencies.

Notice the different symbols on the software launch project's Gantt chart (Figure 10):

The black diamond symbol represents a milestone—a significant event on a project with zero duration. In Figure 10, Task 1, "Marketing Plan distributed," is a milestone that occurs on March 17. Tasks 3, 4, 8, 9, 14, 25, 27, 43, and 45 are also milestones. For large projects, senior managers might want to see only milestones on a Gantt chart.

The thick black bars with arrows at the beginning and end represent summary tasks. For example, activities 12 through 15—"Develop creative briefs," "Develop concepts," "Creative concepts," and "Ad development"—are all subtasks of the summary task called Advertising, Task 11. WBS activities are referred to as tasks and subtasks in most project management software.

The light gray horizontal bars represent the duration of each individual task. For example, the light gray bar for Subtask 5, "Packaging," starts in mid-February and extends until early May.

Arrows connecting these symbols show relationships or dependencies between tasks. Gantt charts often do not show dependencies, which is their major disadvantage. If dependencies have been established in Project 2000 or other project management software, they are automatically displayed on the Gantt chart.

Tracking Gantt charts can also be used to display planned versus actual schedule progress (Schwalbe, 2002). The main advantage of using Gantt charts is that they provide a standard format for displaying project schedule information, and they are easy to create and understand. The main disadvantage of Gantt charts is that they do not usually show relationships or dependencies between tasks. If Gantt charts are created using project management software and tasks are linked, then the dependencies would be displayed, but not as clearly as they would be displayed on project network diagrams.

Project Network Diagrams

A project network diagram is a schematic display of the logical relationships or sequencing of project activities. To use critical path analysis, one must determine task relationships. A critical path for a project is the series of activities that determines the earliest time by which the project can be completed. It is the longest path through the network diagram and has the least amount of slack or

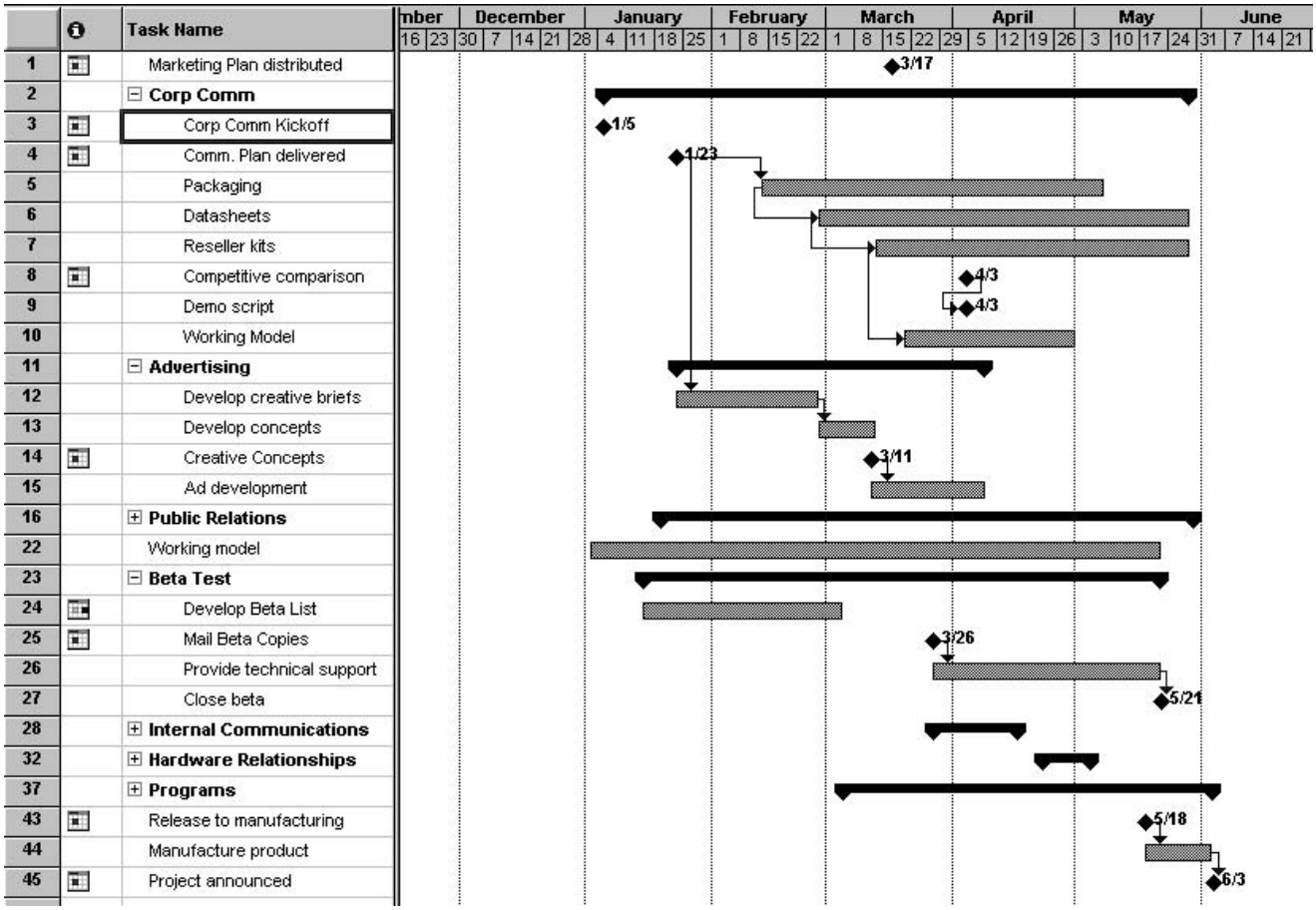


Figure 10: Gantt chart for software launch project in Microsoft Project.

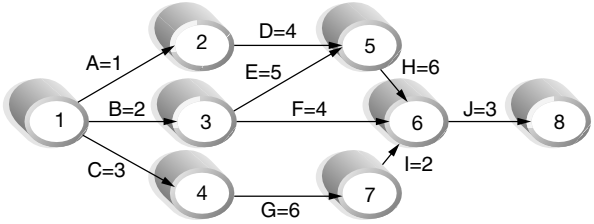
float. Slack or float is the amount of time an activity may be delayed without delaying a succeeding activity or the project finish date.

To find the critical path for a project, a project manager should first develop a good network diagram, which, in turn, requires a good activity list based on the WBS. Once a project network diagram has been created, the manager must also estimate the duration of each activity to determine the critical path. Calculating the critical path involves adding the durations for all activities on each path through the project network diagram. The longest path is the critical path.

Figure 11 provides an example of a project network diagram using the activity-on-arrow format for simplicity. Note that each of the four paths starts at the first node (1) and ends at the last node (8) on this diagram. This figure also shows the length or total duration of each path through the project network diagram. These lengths are computed by adding the durations of each activity on the path. Because path B-E-H-J at 16 days has the longest duration, it is the critical path for the project.

Figure 12 shows the network diagram and critical path when the same project information is entered into Microsoft Project 2000. This format is called the precedence diagramming method (PDM) and is used by most software packages. Project management software can definitely help in creating network diagrams and determining critical paths and slack amounts for all project tasks.

What does the critical path really mean? The critical path shows the shortest time in which a project can be completed. Even though the critical path is the longest path, it represents the shortest time it takes to complete a project. If one or more of the activities on the critical path takes longer than planned,the whole project schedule



Note: Assume all durations are in days.

Path 1:	A-D-H-J	Length = 1+4+6+3 = 14 days
Path 2:	B-E-H-J	Length = 2+5+6+3 = 16 days
Path 3:	B-F-J	Length = 2+4+3 = 9 days
Path 4:	C-G-I-J	Length = 3+6+2+3 = 14 days

Since the critical path is the longest path through the network diagram, Path 2, B-E-H-J, is the critical path for Project X.

Figure 11: Using a project network diagram to determine the critical path.

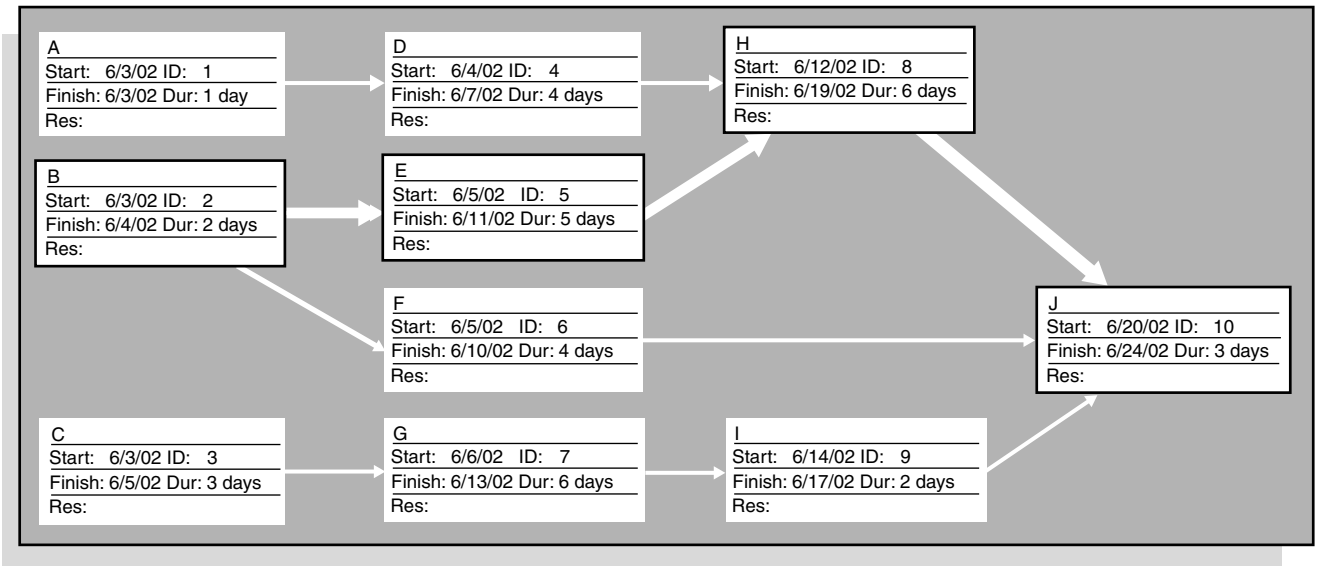


Figure 12: Network diagram and critical path in Project 2000.

will slip unless the project manager takes corrective action.

People are often confused about what the critical path is for a project or what it really means. Some people think the critical path includes the most critical activities, but it is concerned only with the time dimension of a project. Just because its name includes the word critical does not mean that it includes all critical activities. For example, Frank Addeman, executive project director at Walt Disney Imagineering, explained in a keynote address at the May 2000 PMI-ISSIG Professional Development Seminar (http://www.pmi-issig.org) that growing grass was on the critical path for building Disney's Animal Kingdom theme park! This 500-acre park required special grass for its animal inhabitants, and some of the grass took years to grow. Another misconception is that the critical path is the shortest path through the project network diagram. In some areas, for example, transportation modeling, similar diagrams are drawn in which identifying the shortest path is the goal. For a project, however, each activity must be done to complete the project. It is not a matter of choosing the shortest path.

Program Evaluation Review Technique

Another project time management technique is the Program Evaluation and Review Technique (PERT), a network analysis technique used to estimate project duration when there is a high degree of uncertainty about the individual activity duration estimates. PERT applies the critical path method to a weighted average duration estimate.

PERT uses probabilistic time estimates—duration estimates based on using optimistic, most likely, and pessimistic estimates of activity durations—instead of one specific or discrete duration estimate. Like the critical path method, PERT is based on a project network diagram, normally the PDM method. To use the PERT method, one calculates a weighted average for the duration estimate of each project activity using the following formula:

PERT weighted average = optimistic time + 4X most likely time + pessimistic time

The main advantage of PERT is that it attempts to address the risk associated with duration estimates. PERT has three disadvantages: it involves more work because it requires several duration estimates, there are better probabilistic methods for assessing risk (such as Monte Carlo simulation), and it is rarely used in practice. In fact, many people confuse PERT with project network diagrams because the latter are often referred to as PERT charts.

Critical Chain Scheduling

A variation of critical path analysis that addresses the challenge of meeting or beating project finish dates is an application of the theory of constraints called critical chain scheduling. The theory of constraints is based on the fact that, like a chain with its weakest link, any complex system at any point in time, often has only one aspect or constraint that limits its ability to achieve more of its goal. For the system to attain any significant improvements, that constraint must be identified, and the whole system must be managed with it in mind. Critical chain is a method of scheduling that takes limited resources into account when creating a project schedule and includes buffers to protect the project completion date.

One can find the critical path for a project without considering resource allocation. For example, task duration estimates and dependencies can be made without considering the availability of resources. In contrast, an important concept in critical chain scheduling is the availability of resources. If a particular resource is needed full-time to complete two tasks that were originally planned to occur simultaneously, critical chain scheduling acknowledges that one of those tasks must be delayed until the resource is available or another resource must

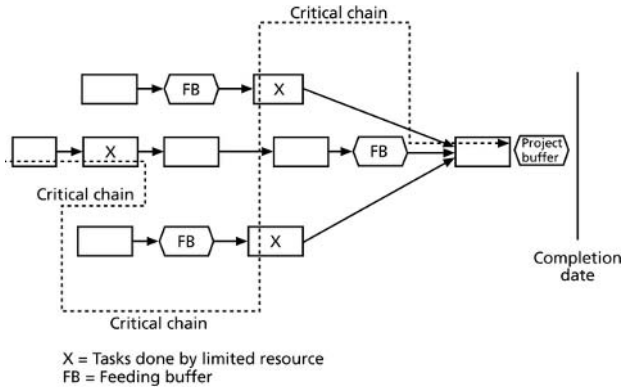


Figure 13: Example of critical chain scheduling.

be found. In this case, accounting for limited resources often lengthens the project finish date, which is not most people's intent.

Other important concepts related to critical chain include multitasking and time buffers. Critical chain scheduling assumes that resources do not multitask. Someone cannot be assigned to two tasks simultaneously on the same project using critical chain scheduling. Likewise, critical chain theory suggests that projects be prioritized so people working on more than one project at a time know which tasks take priority. Preventing multitasking avoids resource conflicts and wasted setup time caused by shifting between multiple tasks over time. Critical chain scheduling also changes the way most people make task duration estimates. People often add a safety or buffer, which is additional time to complete a task added to an estimate to account for various factors. These factors include the negative effects of multitasking, distractions and interruptions, fear that estimates will be reduced, Murphy's Law, and so on. Critical chain scheduling removes buffers from individual tasks and instead creates a project buffer, which is additional time added before the project's due date. Critical chain also protects tasks on the resource-constrained critical path from being delayed by using feeding buffers, which are additional time added before critical path tasks that are preceded by non-critical-path tasks.

Figure 13 provides an example of a project network diagram constructed using critical chain scheduling. Note that the critical chain accounts for a limited resource, X, and the schedule includes use of feeding buffers and a project buffer in the network diagram. The task estimates in critical chain scheduling should be shorter than traditional estimates because they do not include their own buffers. By not having task buffers, there should be less occurrence of Parkinson's Law, which states that work expands to fill the time allowed. The feeding buffers and project buffers protect the date that really needs to be met—the project completion date (Goldratt, 1997).

TECHNIQUES FOR SHORTENING A PROJECT SCHEDULE

It is common for stakeholders to want to shorten a project schedule estimate. By knowing thecritical path, the project manager and his or her team can use several duration compression techniques to shorten the project schedule. One simple technique is to reduce the duration of activities on the critical path. The manager can shorten the duration of critical path activities by allocating more resources to those activities or by changing their scope. Two other techniques for shortening a project schedule are crashing and fast tracking.

Crashing is a technique for making cost and schedule trade-offs to obtain the greatest amount of schedule compression for the least incremental cost. For example, suppose one of the items on the critical path for a project was entering data into a database. If this task is yet to be done and was originally estimated to take 2 weeks based on the organization providing one part-time data-entry clerk, the project manager could suggest that the organization have the clerk work full-time to finish the task in 1 week instead of 2. This change would not cost the organization more money, and it could shorten the project end date by 1 week. By focusing on tasks on the critical path that could be done more quickly for no extra cost or a small cost, the project schedule can be shortened. The main advantage of crashing is shortening the time it takes to finish a project. The main disadvantage of crashing is that it often increases total project costs.

Another technique for shortening a project schedule is fast tracking. Fast tracking involves doing activities in parallel that would normally be done in sequence or in slightly overlapping time frames. For example, a project team may have planned not to start any of the coding for a system until all of the analysis was done. Instead, team members could consider starting some coding activity before all of the analysis is complete. The main advantage of fast tracking, like crashing, is that it can shorten the time it takes to finish a project. The main disadvantage of fast tracking is that it can end up lengthening the project schedule because starting some tasks too soon often increases project risk and results in rework.

PROJECT COST MANAGEMENT AND PERFORMANCE TRACKING TECHNIQUES

Many information technology projects are never initiated because information technology professionals do not understand the importance of knowing basic accounting and finance principles like net present value analysis, return on investment, and payback analysis. Likewise, many projects that are started never finish because of cost management problems. An important topic and one of the key tools and techniques for controlling project costs is earned value management.

Earned value management (EVM) is a project performance measurement technique that integrates scope, time, and cost data. Given a cost performance baseline, project managers and their teams can determine how well the project is meeting scope, time, and cost goals by entering actual information and then comparing it to the baseline. A baseline is the original project plan plus

Table 3 Earned Value Formulas

TERM	FORMULA
Earned Value (EV)	EV = budgeted cost of work performed
Cost Variance (CV)	CV = EV – AC
Schedule Variance (SV)	SV = EV – PV
Cost Performance Index (CPI)	CPI = EV/AC
Schedule Performance Index (SPI)	SPI = EV/PV
Estimate at Completion	Budget at completion (BAC)/CPI
Estimated Time to Complete	Target time estimate/ SPI

Note: From Schwalbe (2002). © 2002 by Course Technology, a division of Thompson Learning. Reprinted with permission.
PV = planned value.

approved changes. Actual information includes whether a WBS item was completed or approximately how much of the work was completed, when the work actually started and ended, and how much it actually cost to do the work that was completed.

Earned value management involves calculating three values for each activity or summary activity from a project's WBS. Table 3 summarizes important earned value formulas. Note that having a good WBS is an important prerequisite for using earned value management, just as it is for critical path analysis.

1. The planned value (PV), formerly call the budgeted cost of work scheduled (BCWS), also called the budget, is that portion of the approved total cost estimate planned to be spent on an activity during a given period.
2. The actual cost (AC), formerly called the actual cost of work performed (ACWP), are the total direct and indirect costs incurred in accomplishing work on an activity during a given period.
3. The earned value (EV), formerly called the budgeted cost of work performed (BCWP), is the value of the physical work actually completed.

Note that in general, negative numbers for cost and schedule variance indicate problems in those areas. Negative numbers mean the project is costing more or taking longer than planned. Likewise, CPI and SPI less than one or less than 100% also indicate problems.

Earned value calculations for all project activities (or summary level activities) are required to estimate the earned value for the entire project. Some activities may be over budget or behind schedule, but others may be under budget and ahead of schedule. By adding all of the earned values for all project activities, one can determine how the project as a whole is performing.

The project manager can graph earned value information to track project performance. Figure 14 shows an earned value chart for a 1-year project after 5 months. Viewing earned value information in chart form helps

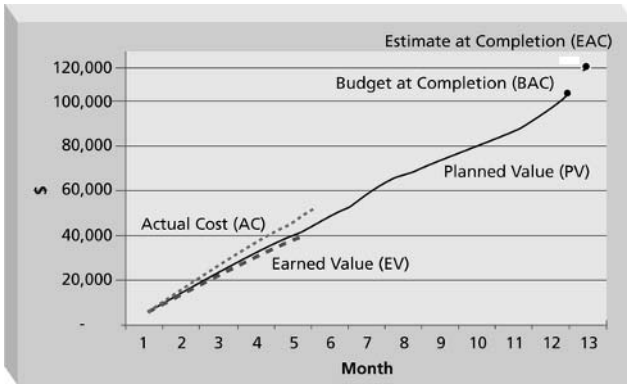


Figure 14: Sample earned value chart.

to visualize how the project is performing. For example, one can see the planned performance by looking at the planned value line. If the project goes as planned, it will finish in 12 months and cost $100,000, represented by the budget at completion (BAC) point. Notice in this example that the actual cost line is always right on or above the earned value line. When the actual cost line is right on or above the earned value line, costs are equal to or more than planned. The planned value line is pretty close to the earned value line, just slightly higher in the last month. This relationship means that the project has been right on schedule until the last month, when the project got a little behind schedule.

Senior managers overseeing multiple projects often like to see performance information in a graphical form like this earned value chart. Earned value charts allow one to see quickly how projects are performing. If there are serious cost and schedule performance problems, senior management may decide to terminate projects or take other corrective action. The estimates at completion are important inputs to budget decisions, especially if total funds are limited. Earned value management is an important technique because, when used effectively, it helps senior management and project managers evaluate progress and make sound management decisions. Many Government projects use earned value, and more private corporations are starting to use this powerful tool (Office of Undersecretary of Defense, n.d.).

OTHER IMPORTANT TOOLS AND TECHNIQUES

As mentioned earlier, there are several project management tools and techniques for each of the nine project management knowledge areas. There are also many tools and techniques used in general management and in application areas that can be applied to projects. This chapter describes just a few of the most common project management tools and techniques. Table 4 lists these and some other common tools and techniques for managing projects. There are many more. Consult PMI (2000b), Schwalbe (2002), or other resources for additional information.

Table 4 Common Tools and Techniques for Managing Projects

KNOWLEDGE AREA/CATEGORY	TOOLS AND TECHNIQUES
Integration	Project management software (Microsoft Project, Primavera, Artemis, Welcom, etc.)
Management	Project plans Change tracking and control techniques Change control boards Configuration management Project review meetings Project leadership Executive sponsorship
Project Selection	Focusing on broad organizational goals Categorizing projects Net present value, return on investment, payback Weighted scoring models Strengths, weaknesses, opportunities, and threats (SWOT) analysis Enterprise project management Multiproject management Project portfolio management
Scope Management	Project charters Scope statements Word breakdown structures Statements of work Requirements analysis Prototyping Extreme programming Joint application design Issue tracking Scope change control
Time Management	Gantt charts Project network diagrams Critical path analysis Program evaluation review technique Critical chain scheduling Crashing Fast tracking Milestone reviews
Cost Management	Earned value management Economic value added Cost estimates (analogous, bottom-up, parametric) Cost Management plan Financial software
Quality Management	Six sigma Quality assurance Quality control charts Pareto diagrams Fishbone or Ishakawa diagrams Quality audits Maturity models Statistical methods
Human Resource Management	Responsibility assignment matrices Resource histograms Resource pools Team building methods (Myers–Briggs-type indicator, social styles profile, physical challenges, etc.) Motivation techniques Self-directed work teams Empathic listening

Table 4 (Contintued)

KNOWLEDGE AREA/CATEGORY	TOOLS AND TECHNIQUES
Communications Management	Stakeholder analysis Communications management plan Conflict management Communications media selection Communications infrastructure Status reports Meetings Virtual communications Templates Project Web sites
Procurement Management	Make or buy analysis Contracts Requests for proposals or quotes Source selection Negotiating E-procurement
Risk Management	Risk management plan Probability impact matrix Risk ranking Monte Carlo simulation Top-Ten Risk Item Tracking

CONCLUSION

Despite all the uncertainty in the world, one can be certain that there will continue to be a need for projects and better ways to manage them. Many organizations have improved project success rates by applying some standard project management processes and using appropriate tools and techniques. This chapter summarizes a few of the common tools and techniques used in project management. Because every project is unique, project managers and their teams must have a good understanding of what tools and techniques are available before they can make the more difficult decisions of which ones to use on their projects and how to implement them.

ACKNOWLEDGMENT

Figures, tables, and most of the text in this article are taken from the author's text, *Information Technology Project Management*, (2nd ed.). Boston: Course Technology, 2002. They are reprinted here with permission of the publisher.

GLOSSARY

Activity or task An element of work, normally found on the work breakdown structure, that has an expected duration, cost, and resource requirements.

Baseline The original project plan plus approved changes.

Crashing A technique for making cost and schedule trade-offs to obtain the greatest amount of schedule compression for the least incremental cost.

Critical chain scheduling A method of scheduling that takes limited resources into account when creating a project schedule and includes buffers to protect the project completion date.

Critical path The series of activities that determines the earliest time by which the project can be completed.

Earned value management A project performance measurement technique that integrates scope, time, and cost data.

Fast tracking Compressing a schedule by doing activities in parallel that you would normally do in sequence or in slightly overlapping time frames.

Gantt chart A standard format for displaying project schedule information by listing project activities and their corresponding start and finish dates in a calendar format.

Knowledge areas Topics that describe key competencies project managers must develop to manage projects effectively. (The nine knowledge areas in project management are project integration, scope, time, cost, quality, human resource, communications, risk, and procurement management.)

Milestone A significant event on a project with zero duration.

Organizational breakdown structure A structure that describes the people responsible for performing project work.

Program evaluation review technique A network analysis technique used to estimate project duration when there is a high degree of uncertainty about the individual activity duration estimates.

Project charter A document that formally recognizes the existence of a project and provides direction on the project's objectives and management.

Project management The application of knowledge, skills, tools, and techniques to project activities in order to meet project requirements.

Project network diagram A schematic display of the logical relationships or sequencing of project activities.

Project A temporary endeavor undertaken to accomplish a unique purpose.

Responsibility assignment matrix A matrix that maps the work of the project to the people responsible for performing the work.

Slack or float The amount of time an activity may be delayed without delaying a succeeding activity or the project finish date.

Stakeholders The people involved in or affected by project activities.

Statement of Work A form of a scope statement often used by the government or in contracts.

Triple constraint A project's scope, time, and cost goals or constraints.

Weighted scoring model A tool that provides a systematic process for selecting projects based on many criteria.

Work breakdown structure A deliverable-oriented grouping of the work involved in a project that defines the total scope of the project.

Work package A deliverable or product at the lowest level in a work breakdown structure.

CROSS REFERENCES

See *Electronic Commerce and Electronic Business; Prototyping; Return on Investment Analysis for E-business Projects; Software Design and Implementation in the Web Environment*.

REFERENCES

Goldratt, E. M. (1997). *Critical chain*. Great Barrington, MA: North River Press. (Also available at http://www.goldratt.com)

Office of the Under Secretary of Defense (n.d.). Acquisition resources & analysis/acquisition management. Earned value management. Retrieved April 19, 2002, from http://www.acq.osd.mil/pm

Project Management Institute (2001a) *The PMI project management fact book* (2nd ed.). Newtown Square, PA: Author.

Project Management Institute (2001b). *Practice standard for work breakdown structures*. Newtown Square, Pennsylvania: Project Management Institute. (Also available at http://www.pmi.org/standards)

Project Management Institute (2000a). *PMI's 2000 salary survey*. Newtown Square, PA: Author.

Project Management Institute (2000b). *PMBOK Guide, 2000 edition*. Newtown Square, PA: Author.

Schwalbe, K. (2002). Information technology project management (2nd ed.). Boston: Course Technology. (Readers can access lecture notes for this text and many references related to project management from the author's Web site at http://www.augsburg.edu/ppages/~schwalbe)

Stewart, T. A., & McGowan, J. (1996, March 20). Planning a career in a world without managers. *Fortune*.

The Standish Group (2001). *CHAOS 2001: A recipe for success*. West Yarmouth, MA: Author.

FURTHER READING

For links to many project management software products and general information on project management: http://www.allpm.com/ (Retrieved April 19, 2002).

For links to many project management software products: http://www.infogoal.com/pmc/pmcswr.htm (Retrieved April 19, 2002).

For definitions of several financial terms: http://www.investopedia.com/dictionary (Retrieved: April 19, 2002).

Propagation Characteristics of Wireless Channels

P. M. Shankar, *Drexel University*

INTRODUCTION

Wireless systems encompass a wide range of information transmission mechanisms, including cordless phones, paging systems, cell phones, satellite communication systems, maritime-mobile systems, industrial and medical monitoring systems, infrared (IR) remote controls, and so on. These systems have unique operating frequency bands. The choice of the frequencies is often determined by the range of operation (a few meters for a cordless phone to thousands of kilometers for satellite communication systems), the medium through which the signal traverses (e.g., urban areas with tall structures, vast empty spaces in rural areas, and multilevel, multistructure environments in factories and malls), the amount of data to be transmitted (low volume of data in maritime mobile systems and IR-based remote control systems to Gbits/s in satellite systems). The other factors that are equally important are the size of the transmitting and receiving antennas and the cost and maintenance of installations. A broad classification of various frequencies used in wireless communications is shown in Table 1; abbreviation in table are used subsequently in text.

Two important considerations of communication systems must be kept in mind:

1. The physical dimensions of an antenna are inversely proportional to its frequency. Thus, at low frequencies, the size of the antenna may be too large for these frequencies to be used in mobile systems. For example, at 30 kHz, the wavelength is 10,000 m:

$$\left[\text{wavelength} = \frac{\text{velocity}}{\text{frequency}} = \frac{3 \times 10^8}{30 \times 10^3} = 10{,}000 \text{ m}\right].$$

   Note that the required sizes of the antennas will be on the order of a half wavelength. This means that for transmission at 30 kHz, we would require an antenna of length 5,000 m.
2. The amount of data that can be transmitted is directly proportional to the frequency. The data rates at 30 kHz will be less than 30 kbps, whereas the data rates at 1 GHz can be on the order of 1 Gbps. The actual data rates will be determined by a number of factors such as the transmission distance, the characteristics of the intervening medium, and the modulation–demodulation schemes.

Once we take these considerations into account, it is clear that the very low frequencies can be eliminated from the range of operation of the practical wireless communication systems, which demand transmission at very high data rates, portability (i.e., mobile systems), and compact-size. The large size of antennas and low data rates associated with low frequencies eliminate them from considerations in modern wireless communication systems and networks. Even though the frequencies on the higher end of the spectrum (Table 1) appear to offer smaller size antennae and higher data rates, their use in wireless systems is limited by other factors. The primary reason for their unsuitability is that these frequencies are severely attenuated by "obstructions" in their path. A case of a direct path (line of sight [LOS]) between the transmitter and a receiver is shown in Figure 1.

In a typical urban environment containing tall buildings and other structures, these signals (EHF) will be unable to penetrate and reach the transmitter (Bertoni, Honcharenko, Maciel, & Xia, 1994; Jakes, 1974; Pardo, Cernicharo, & Serabyn, 2001; Parsons, 1996; Saunders, 1999; Steele & Hanzo, 1999). They require a clear line of sight between the transmitter and receiver, making them well suited for satellite-to-ground (or ground-to-satellite) transmission or satellite-to-satellite transmission. They may also be used in very specific applications involving short-range transmission of a large volume of data in confined spaces with no obstructions. Thus, it is possible to narrow down the choice of the frequencies for practical wireless systems based on the additional consideration of the manner in which the signals reach the receiver from the transmitter. Frequencies that can penetrate buildings, be reflected, refracted, diffracted, and scattered from buildings, trees, and still reach the receiver must be available. The important considerations for the choice of frequencies can now be restated as follows:

1. Size of the antenna
2. Ability and convenience to mount the antenna on portable units, moving vehicles, and so on
3. Ability to reach the receiver even when a LOS path is not available
4. Data rates

These important factors narrow the frequencies to UHF or typically in the range of 900 MHz to 3 GHz. Additional factors that affect the performance of the wireless systems will be discussed later.

Table 1 Various Frequency Bands Used in Wireless Systems, Their Identification, and Their Characteristics

FREQUENCY BAND	FREQUENCY RANGE	ANTENNA SIZE	DATA RATE
VLF (very low frequency)	3–30 kHz	Large	Low
Low frequency (LF)	30–300 kHz	↑	↓
Medium frequency (MF)	300 kHz–30 MHz		
High frequency (HF)	3–30 MHz		
Very high frequency (VHF)	30–300 MHz		
Ultra high frequency (UHF)	300 MHz–3 GHz		
Super high frequency (SHF)	3–30 GHz		
Extra high frequency (EHF)	30–300 GHz	Small	High

Having narrowed the operational frequency bands in practical wireless systems, we now look at two key factors that have a significant impact on the performance of wireless communication systems: the signal strength received and any variability of the signal strength. These factors play a major role in determining how far the signal will travel (i.e., separation between transmitter and receiver) and how much data can be transmitted usefully. The former is dictated by the level of loss suffered by the signal as it reaches the receiver, and the latter is dictated by the level variations. Both are uniquely determined by the terrain between the transmitter and receiver.

PROPAGATION OF SIGNALS

The following sections consider the attenuation and the signal variability experienced by the wireless signals.

Transmission Loss

As the electromagnetic waves travel from the transmitter to the receiver, they encounter various objects in their path. In typical urban environments, it may not be possible to have an LOS path between the transmitter and the receiver. The signal leaving the transmitter reaches the receiver through a number of mechanisms, such as reflection, diffraction, and scattering (Feher, 1995; IEEE, 1988; Jakes, 1974; Rappaport, 2002; Shankar, 2001). These are depicted in Figure 2. Reflection occurs when the signal encounters objects that are much larger than the wavelength. This is shown in Figure 2 (left). While reflection is taking place, refraction of the wave may also take place, in which case the signal will penetrate the object, which may be a wall or a partition. The signal may also undergo diffraction (i.e., bending over obstacles) when the signal encounters sharp boundaries as seen in Figure 2 (center). Scattering occurs when the surface of the object is rough and is on the order of the wavelength. The signal is scattered in all directions as seen in Figure 2 (right) as opposed to getting reflected in a specific direction in the case of reflection where the angle of incidence is equal to the angle of reflection. Thus, in outdoor areas, a signal reaches the receiver through reflection, scattering, and diffraction, as it encounters building, trees, and other artificial or natural objects. In indoor areas, such as in malls, factories, and office buildings, the signal will reach the receiver penetrating floors, walls, ceilings, and so on while undergoing effects of reflection, scattering, and diffraction.

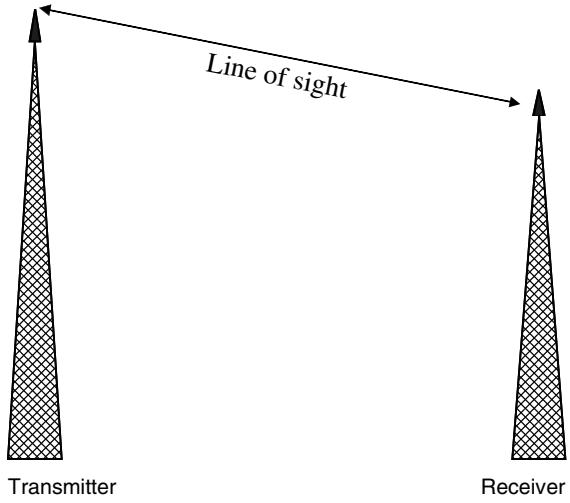


Figure 1: A line of sight propagation is shown.

A typical signal received in urban wireless systems is shown in Figure 3a. It is seen that the received power decreases as the distance increases. The power is plotted in dBm. (The relationship between power in watts and dBm is given in the Appendix.) In a short segment of this curve, as shown in Figure 3b, the power loss is not a straight line. The power fluctuates as it decreases. These fluctuations are referred to as long-term fading or shadowing (Pahlavan & Levesque, 1993; Parsons, 1996). Zooming further into the power versus distance curve, the power fluctuates around a mean value (Fig. 3c). These fluctuations are of short duration compared with those seen in Figure 3b and are referred to as short-term fading. They are also known as Rayleigh fading, based on the statistical fluctuations in the received envelope of the signal (Rappaport, 2002; Shankar, 2001; Steele & Hanzo, 1999; Stein, 1987). Nakagami fading and Rician fading are also used to describe short-term fading.

It is thus clear that the signal undergoes attenuation and fluctuations as it reaches the receiver. The fluctuations introduce random variations in the received power, making it necessary to take additional steps to design the link. The next section explores ways of modeling the power loss and the fluctuations in power.

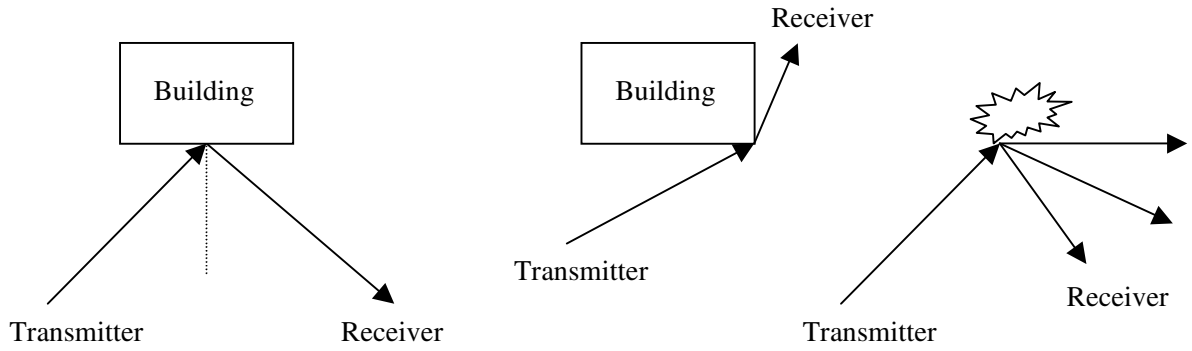


Figure 2: Various propagation mechanisms are shown: reflection (left), diffraction (center), and scattering (right).

Modeling of Power Loss

Several models are available to predict the median value of the received power (Har, Xia, & Bertoni, 1999; Hata, 1980; Lee, 1986; Oda, Tsunkewa, & Hata, 2000; Okumura, Ohmori, & Fukuda, 1968; Vogel & Hong, 1988). These are available for propagation outdoors and indoors (Bultitude, Mahmoud, & Sullivan, 1989; Durgin, Rappaport, & Xu, 1998; Harley, 1989; Lott & Forkel, 2001; Rappaport & Sandhu, 1994). They are also available for various frequency bands that are of interest in wireless communications. Instead of concentrating on these models, I initially describe one of the simple ways of predicting loss based on the concept of the *path loss exponent*. To understand this concept, consider the case of a line-of-sight propagation in free space as shown in Figure 1. If P_t is the transmitted power in W (Watts), the received power P_d (W) at a distance d from the transmitter is given by Friis formula (IEEE, 1988),

$$P_d = P_t G_t G_r \left(\frac{\lambda}{4\pi d}\right)^2 W, \tag{1}$$

where G_t = gain of the transmitting antenna, G_r = gain of the receiving antenna, and λ = free space wavelength = $\frac{c}{f_0}$,

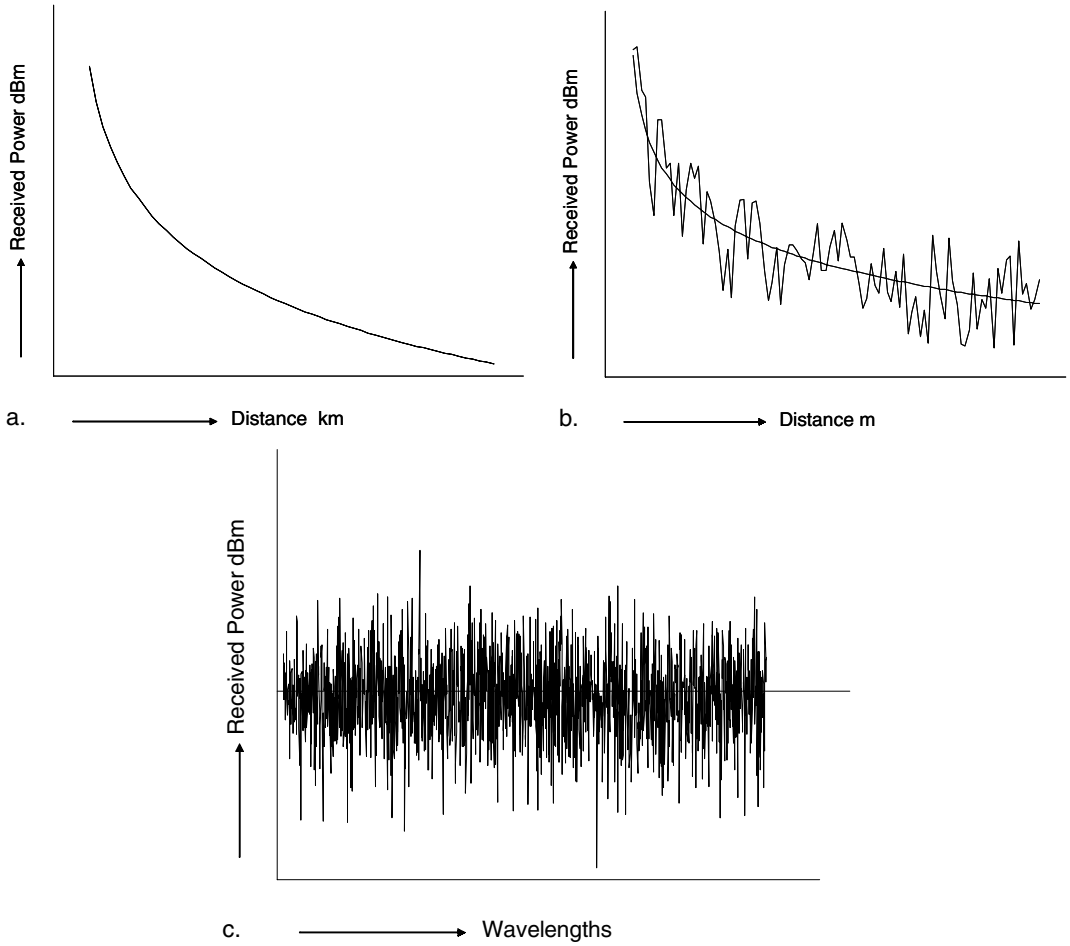


Figure 3: The plot of the received signal: (a) attenuation; (b) long-term fading; (c) short-term fading.

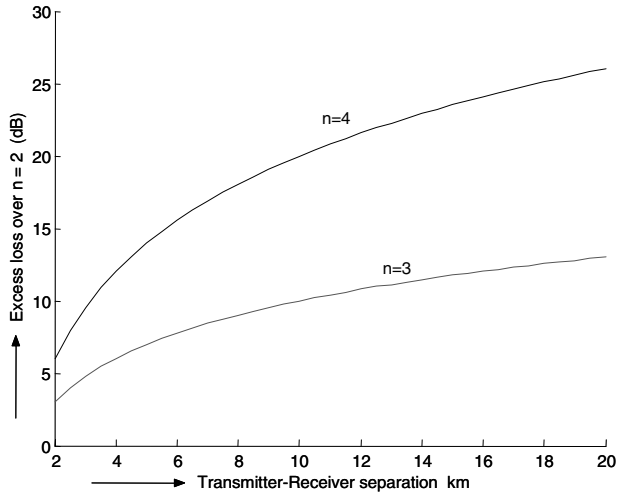


Figure 4: The losses at $n = 3$ and $n = 4$ are compared with the free space loss ($n = 2$).

c being the velocity of light and f_0 the frequency. Note that G_t, and G_r, are dimensionless (i.e., numbers) and d and λ must have the same units (centimeters, meters, or kilometers). Assuming equal gain antennas ($G_t = G_r$), the received power can be expressed as inversely proportional to the square of the distance

$$P_d \propto \frac{1}{d^2}.$$

Conversely, we can say that the loss experienced by the signal is directly proportional to the square of the distance. The path loss exponent or the path loss coefficient determines the decay of the power as distance increases and is denoted by n. In free space under LOS conditions, the path loss exponent n is 2. Because there are no obstacles in the path of the signal in LOS propagation in free space, no reflection, diffraction, or scattering takes place, $n = 2$ will be the best case scenario expected in signal transmission. In a general case, where propagation takes place in a region containing obstacles, the path loss exponent will be larger than 2, pointing to higher path loss as n increases. The excess loss over $n = 2$ is plotted in Figure 4 as a function of the distance. The loss at a given distance increases as n goes up. Note that the path loss exponent n is also sometimes referred to as power decay index, distance power gradient, or slope factor.

It must be noted that the low values of n also increase cochannel interference, namely, the interference coming from other cells using the same channel. As the value of n increases, the interference goes down, leading to an improvement in the capacity of the cellular communication systems.

Calculation of the Received Power at Any Distance

Power received at any distance d is inversely proportional to the nth power of the distance,

$$P_d \propto \frac{1}{d^n}. \tag{2}$$

Equation 2 cannot be applied directly because of the need to evaluate the proportionality factor. This is done by applying Frii's equation (Feher, 1995) at a very short distance d_0 from the transmitter, where it can be assumed that free space LOS conditions exist. If P_{d_0} is the power at a distance of d_0 from the transmitter ($G_t = G_r = 1$),

$$P_{d_0} = P_t \left(\frac{\lambda}{4\pi d_0}\right)^2. \tag{3}$$

Using Equation 2 one can now write

$$\frac{P_d}{P_{d_0}} = \frac{d_0^n}{d^n}. \tag{4}$$

Taking the logarithm, one arrives the expression for the received power at a distance d ($d > d_0$) from the transmitter to be

$$P_d(dBm) = P_{d_0}(dBm) - 10n\log_{10}\left(\frac{d}{d_0}\right), \tag{5}$$

where

$$P_{d_0}(dBm) = P_t(dBm) + 20\log_{10}\left(\frac{\lambda}{4\pi d_0}\right). \tag{6}$$

Note that in Equations 5 and 6, the power is expressed in decibel units (dBm). The question is now what value of d_0 is appropriate. Typically, this value, known as the *reference distance* ($\lambda < d_0$), is chosen to be 100 m in outdoor environments and 1 m in indoor environments (Pahlavan & Levesque, 1995; Rappaport, 2002; Shankar, 2001).

The models described thus far can be applied for calculating the received power at various operating frequencies indoors and outdoors. A few points are in order. The received power according to Frii's equation (Equation 3) decreases as the wavelength decreases. This means that as one moves from 900 MHz band to the PCS (personal communication systems) operating at the 1,800 to 2,000-MHz band, the received power decreases. As wireless communication systems move into the 4–6 GHz band, the received power decreases further as the wavelength decreases (Table 1). Another factor that becomes critical as the frequencies increase is the inability of the signal to penetrate buildings. For example, compared with 900-MHz signals, 2-GHz signals will not travel far from the transmitter. Signals may even be blocked by a single building between the transmitter and receiver. Thus, the LOS propagation becomes the predominant means by which the signal reaches the receiver. In the case of a microwave signal, a truck obstructing the path can bring down the received signal to extremely small levels (IEEE, 1988).

The approaches based on the path loss exponent are not the only ones to estimate the loss suffered by the signal as it reaches the receiver. The disadvantage of the path-loss-based approach is that it does not directly take into account a number of system dependent factors such the heights of the transmitting and receiving antennas and their locations. Models such as Lee's (1986) model, Hata's

(1980) model, and the Walfish and Bertoni model (Har et al., 1999; Ikegami, Tekeuchi, & Yoshida, 1991; Oda et al., 2000; Okumura et al., 1968; Vogel & Hong, 1988) are also available to calculate the received power. These models take into account the antenna heights and other factors. Yet another approach is to use a two exponent model, where instead of using a single exponent, two values, n_1 and n_2, are used. The loss increases slowly (i.e., the loss occurs as if we have propagation in free space with $n = n_1 = 2$) until a break point is reached. Beyond that point, the loss goes up at a higher rate, with n taking value of n_2 in the range 3 to 9 (Bertoni et al., 1994; Rappaport & Sandhu, 1994). Even though this double exponent approach may appear to lead to better results, some of the path loss prediction models can be used to get more satisfactory results. I now describe the Hata (1980) model for the calculation of the loss suffered by the signal as it travels from the transmitter to the receiver. The loss L_p in decibels (dB) is

$$L_p = P_t(dBm) - P_d(dBm), \tag{7}$$

where the transmitted power P_t and the received power P_d are once again expressed in decibel units.

Hata Model

Based on measurements in a number of cities, Hata (1980) proposed a model to predict the median loss suffered by the wireless signal. This model is an improvement over that proposed by Okumura et al. (1968) because it incorporates correction factors to account for the antennae height based on the geographical location. When one moves from large cities to rural areas, the reduction in power loss can be accommodated with correction factors for the antenna height. The median loss in urban areas according to Hata model, L_p is given by

$$\begin{aligned} L_p(dB) \;=\; & 69.55 + 26.16\log_{10}(f_0) \\ & + (44.9 - 6.55\log_{10} h_b)\log_{10} d \\ & - 13.82\log_{10} h_b - a(h_{mu}), \end{aligned} \tag{8}$$

where f_0 is the carrier frequency (MHz), $400 \le f_0 \le 1500$; d is the separation between base station and mobile unit (km), $d > 1$; h_b is the height of the base station antenna (m); h_{mu} is the height of the mobile unit antenna (m); and $a(h_{mu})$ is the correction factor for mobile unit antenna height.

For large cities, the correction factor $a(h_{mu})$ is given as

$$a(h_{mu}) = 3.2[\log_{10}(11.75\,h_{mu})]^2 - 4.97 \quad (f_0 \ge 400\,\text{MHz}). \tag{9}$$

For small and medium cities, the correction factor is

$$a(h_{mu}) = [1.1\log_{10}(f_0) - 0.7]h_{mu} - [1.56\log_{10}(f_0) - 0.8]. \tag{10}$$

For suburban areas, the median loss, L_{sub} is given by

$$L_{sub}(dB) = L_p - 2\left[\log_{10}\left(\frac{f_0}{28}\right)\right]^2 - 5.4, \tag{11}$$

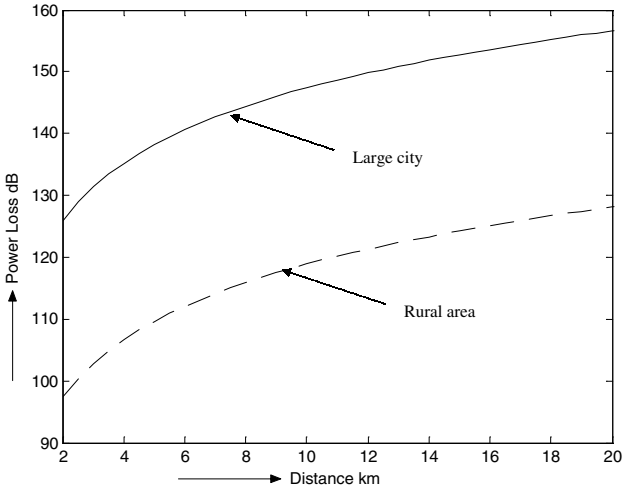


Figure 5: Loss predictions based on Hata model are shown. The radio frequency signal is at 900 MHz. The base station antenna height is 150 m, and the mobile unit antenna height is 1.5 m.

where L_p is the median loss in small–medium cities. For rural areas, the median loss, L_{rur} is given by

$$L_{rur}(dB) = L_p - 4.78[\log_{10}(f_0)]^2 + 18.33\log_{10} f_0 - 40.94. \tag{12}$$

The loss observed in large cities and rural areas is shown in Figure 5. It is easily seen that the losses are higher in urban areas.

The Hata model is not capable of predicting losses in the frequency bands used in some of the present-day systems operating in the 1,500–2,000 MHz band. The Hata model can be extended to cover this range (Rappaport, 2002; Saunders, 1999; Shankar, 2001) and the median power loss in urban areas, $L_P(dB)$ can be expressed as

$$\begin{aligned} L_P(dB) = {} & 46.3 + 33.93\log_{10}(f_0) - 13.82\log_{10}(h_b) \\ & - a(h_{mu}) + [44.9 - 6.55\log_{10}(h_b)]\log_{10} d \\ & + Corr \end{aligned} \tag{13}$$

where $Corr$ is the additional correction factor given by

$$Corr = \begin{cases} 0 & dB \text{ for medium city and suburban areas} \\ 3 & dB \text{ for metropolitan areas} \end{cases}$$

This model is valid for the following parameters only:

f_0 : 1,500–2,000 MHz
h_b : 30–200 m
h_{mu} : 1–10 m
d : 1–20 km

This model is applicable only for distances beyond 1 Km and thus is not applicable in microcells and picocells in which the distances between the transmitter and receiver may be only a few hundred meters. Other models

are available for loss prediction over short ranges (Har et al., 1999; Harley, 1989; Ikegami et al., 1991).

Indoor Wireless Systems

The loss calculation in indoor wireless systems is less straightforward than outdoor systems. It is possible to use Equations 5 and 6 to calculate the received power or predict the loss. Based on empirical measurements conducted indoors and outdoors, the range of values of n has been proposed by several researchers. Note that these values depend on the environments in which the wireless signal is propagating, and in indoor propagation, the values of n are strongly influenced by factors such as the building materials used, floor arrangement, location of the transmitting antenna (inside the building or outside the building), height of the transmitting antenna, frequency used. The values of n range from 2 to 4 as one moves from an open space where free space LOS propagation is possible to urban areas with tall buildings and other structures. In indoor environments, the value of n less than 2 has been observed in grocery stores and open-plan factories (Dersch & Zollinger, 1994; Durgin et al., 1998; Rappaport & Sandhu, 1994). This low value (lower than $n = 2$ in free space) has been attributed to the strong reflections contributed to the received signal by the metallic structures in those places, resulting in higher power levels compared with a completely open space ($n = 2$). Inside the buildings, n can take values in the range of 1.5–4 depending on the number of floors, location of the transmitter, and type of partition (hard vs. soft) used.

A typical indoor propagation scenario is shown in Figure 6. The base station is outside the building. It is also possible to have the base station inside the building depending on the complex. It is easy to see that the signal may have to travel through multiple floors and multiple walls, each of which may be of a different type, to reach the receiver. The loss calculations are therefore complicated, and the wireless systems would have to be designed specifically for a given complex or building. It is nonetheless possible to write a general equation for the calculation of the loss. The loss at a distance d, L_p can be expressed as

$$L_p(d)dB = L_p(d_0)dB + 10n_{eff}\log_{10}\left(\frac{d}{d_0}\right) + \sum_{k=1}^{K} P_k, \qquad (14)$$

where n_{eff} is the effective loss exponent taking into account the multiple floors and walls in the building through which the signal traverses (Durgin et al., 1998; Rappaport and Sandhu, 1994). The other factor in Equation 14, P_k, is the specific material attenuation in dB suffered by the signal as it traverses through K floors/walls. The loss at a distance of $d_0 = 1$ m is given by $L_p(d_0)$.

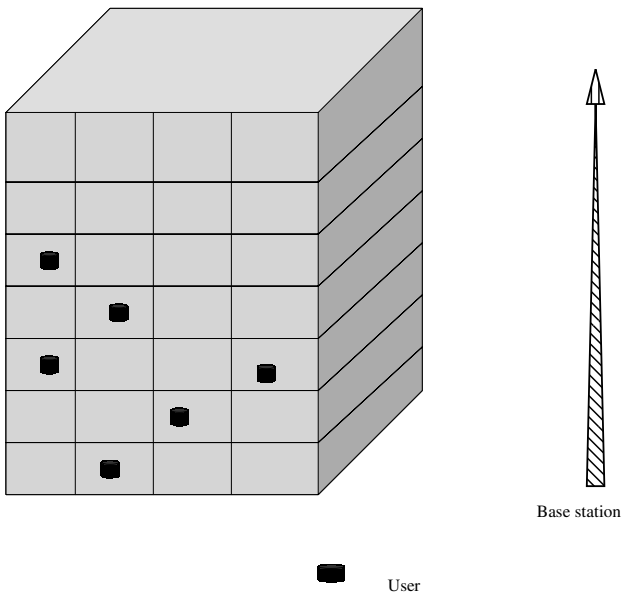


Figure 6: An indoor wireless system with the base station (BS) serving all the users inside. The BS may also be inside the building. The signals between the BS and the users may have to travel through several floors and partitions.

Signal Variability and Fading

I now briefly review the origins of signal variability seen in wireless systems. The signal variability may be caused by short-or long-term fading, or both. Short-term fading may result from multipath fading and Doppler fading. The next sections explores these fading mechanisms and diversity techniques used to mitigate the problems caused by fading.

Multipath Fading

The second characteristic of the wireless signal is the signal variability seen in wireless systems (Feher, 1995; Hashemi, 1993; Kennedy, 1969; Pahlavan & Levesque, 1995; Rappaport, 2002). The signal variability is the lack of predictability of the loss or received power. We had seen in Figure 3 that the power loss fluctuates as the distance increases, with the mean or median fluctuations obeying the nth power of the distance. This random nature of the wireless signals is termed *fading*. This can lead to occasional loss of the signal and network break down because the systems require a minimum amount of power (threshold) to perform satisfactorily and power fluctuations may bring the power below this threshold. This can be taken into consideration by providing a power margin.

Even though the obvious effect of the fading is the random fluctuations of the received power, fading also is responsible for limiting the bandwidth capability of the wireless systems. I first attempt to explain the reasons for the fluctuation in power and effects of fluctuations on data transmission and then explain why the fading limits the bandwidth.

In a typical wireless environment, the signal reaches the receiver after being reflected, scattered, diffracted, or refracted from a number of objects in its path (Jakes, 1974; Shankar, 2001). Thus, the signal does not take a single path to reach the receiver. Instead, the signal reaches the receiver through multiple paths, as shown in Figure 7.

These signals with various amplitudes and phases combine at the receiver. This multipath phenomenon is responsible for the fluctuations in the signal power observed in Figure 3c. This fluctuation in power or fading is short-term fading, of which there are two major consequences. First, the random nature of these fluctuations

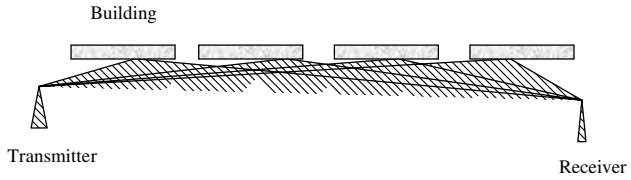


Figure 7: The existence of multiple paths between the transmitter and receiver.

(some times termed *Rayleigh fading*) increases the uncertainty in the received signal power, making it necessary to develop methods to mitigate fading through diversity. Second, the paths shown in Figure 7 take different times leading to a broadening of the received pulse as shown in Figure 8. Figure 8a shows a transmitted pulse $p(t)$. This pulse takes multiple paths and the received pulse $r(t)$ can be expressed as

$$r(t) = \sum_{k=1}^{N} a_k p(t - t_k), \tag{15}$$

where a_k is the strength of the pulse and t_k the time taken by the k^{th} pulse. These delayed pulses of different strengths overlap, broadening the pulse at the receiver seen in Figure 8b. Note that the data rate is inversely proportional to the pulse duration and any broadening of the pulse will lead to overlapping of adjoining pulses resulting in inter symbol interference (ISI). ISI increases the bit error rate (BER). To prevent pulse broadening, it becomes necessary to operate at a lower data rate when fading is present. If pulse broadening leads to a reduction in data rate or makes it necessary to put in place additional signal processing methods to mitigate the effects of pulse broadening, the medium or channel in which this takes place is referred to as *a frequency selective fading channel*. On the other hand, if the pulse broadening is negligible, the medium or channel is referred to as a *flat fading channel*.

Regardless of whether the channel is flat or frequency selective, the fluctuations of power are associated with fading. The fluctuations in power cause an increase in the BER, making it necessary to operate at higher powers. This case is illustrated in Figure 9, which shows the BERs when no fading is present and also when Rayleigh fading is present. To maintain a BER of 1 in 1,000 would require an additional signal-to-noise ratio of approximately 17 dB when fading is present, demonstrating the problems associated with fading.

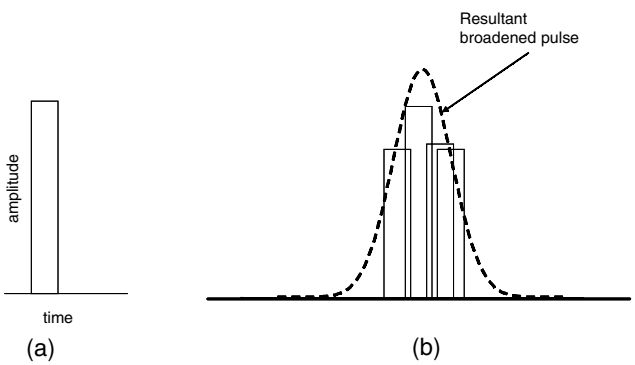


Figure 8: Transmitted pulse (a) and the received pulse (b).

Now consider a signal-to-noise ratio of 10 dB. The results shown in Figure 9 indicate that in the presence of fading a signal-to-noise ratio of 10 dB is not sufficient (Shankar, 2001) to have an acceptable level of performance (say, 1 in 1,000). This leads to outage.

It is possible to have a direct path (LOS) between the transmitter and the receiver in addition to the multiple paths. This condition is more ideal than a pure multipath scenario because the LOS component provides a steady component to the signal. As the strength of the steady component increases (over the multipath components), the deleterious effects of fading decrease. This fading channel is known as the *Rician channel*. It can be shown that as the strength of the LOS component increases, the Rician channel starts approaching the ideal *Gaussian channel,* thus reducing the severity of fading.

Doppler Fading

Multipath fading does not take into account any relative motion of the transmitter and receiver. If the wireless receiver (or transmitter) is mounted on a moving vehicle, the motion introduces Doppler broadening. Motion-induced fading is referred to as *Doppler fading* and is responsible for further degradation in the performance of the wireless systems. If a pure tone of frequency f_0 is transmitted, the received signal spectrum will broaden (Doppler broadening) and contain spectral components raging from $f_0 - f_d$ to $f_0 + f_d$, where f_d is the Doppler shift. If the bandwidth occupied by the information is much greater than the spectral broadening, Doppler spread does not lead to any significant problems in transmission and reception. The channel in this case is referred to as a *slow channel*. If the bandwidth is smaller than the Doppler spread, motion leads to channel varying rapidly within the duration of the pulse. In this case, the channel is referred to as a *fast channel*. Thus, multipath fading decides whether the channel is flat or frequency selective, and Doppler fading decides whether the channel slow or fast (Jakes, 1974; Parsons, 1996). Whereas multipath fading leads to pulse spreading (time), Doppler fading leads to frequency spreading. The different types of fading are summarized in Figure 10.

Long-Term Fading

As shown in Figure 3, power fluctuations have a longer period in Figure 3b than the fluctuations in Figure 3c. These fluctuations with larger periods are statistically described in terms of lognormal fading (Jakes, 1974). Whereas short-term fading is caused by the existence of multipath, long-term fading is caused by the existence of multiple reflections. This means that the signal in a single path cannot reach the transmitter after a single reflection or scattering process but must travel through several structures, causing a "shadowing effect" caused by the presence of many tall structures. The power expressed in decibels in this case is normally distributed, and the power in watts will therefore be lognormally distributed (Jakes, 1974; Rappaport, 2002; Steele & Hanzo, 1999). These

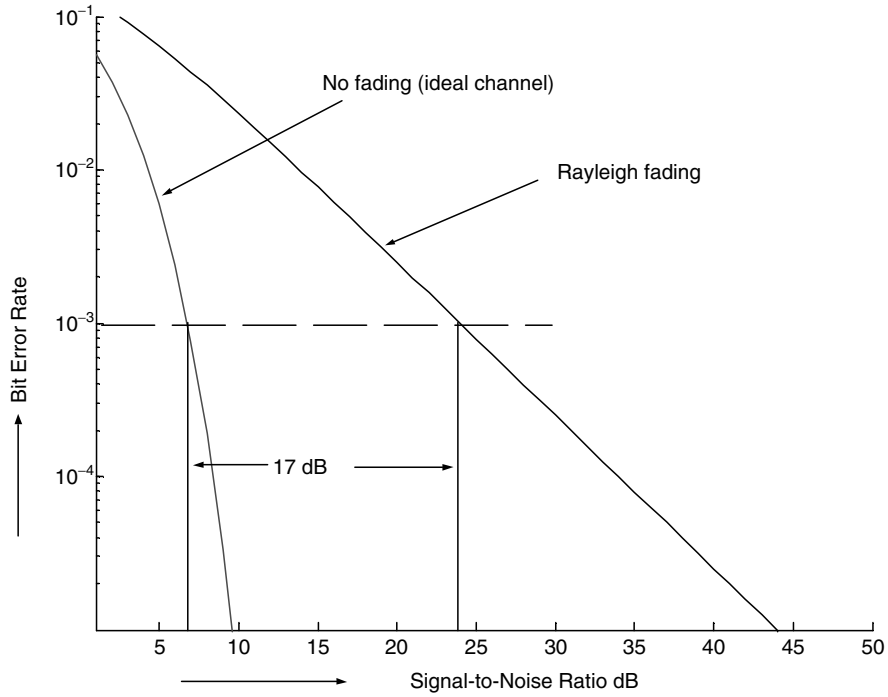


Figure 9: The bit error rate (BER) in an ideal channel and a fading channel. The increased signal-to-noise ratio required to maintain a BER of 1 in 1,000 when fading is present is indicated (17 dB).

situations arise even in indoor channels. To account for the lognormal fading or shadowing, the loss terms in Equation 7 are modified by introducing a term X to the equation for the loss. X is a zero mean normal random variable with a standard deviation of σ_{dB} to take lognormal fading into consideration.

$$Loss = L_p(d) + X \tag{16}$$

The loss calculated now becomes the average of a normally distributed random variable with a standard deviation determined by the severity of the lognormal fading. Long-term fading will also cause outage if the variation in loss is not taken into account during the design of the wireless systems. This situation is handled by including a power margin in the link budget calculations.

Diversity Techniques

Effects of short-term fading can be mitigated through diversity techniques that exploit the randomness existing in the channel. Consider a scenario for which it is possible to

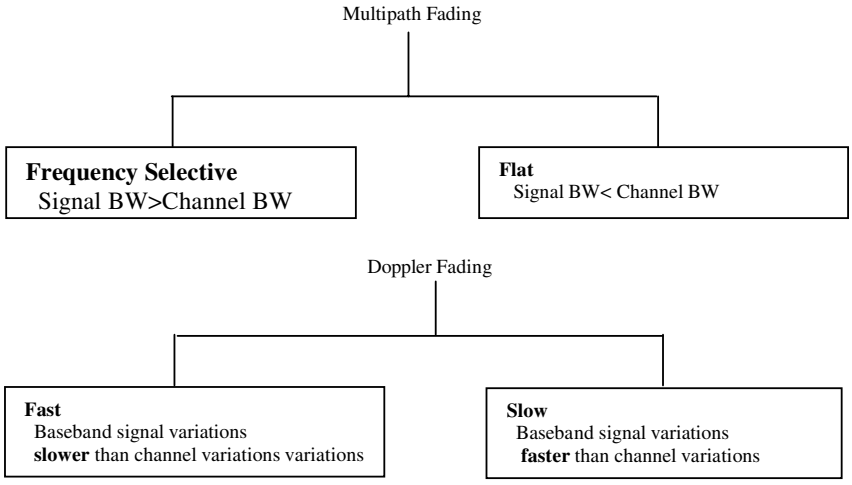


Figure 10: Overview of short-term fading. Pulse broadening taking place in frequency selective channel can be described in terms of the relationship between the message bandwidth and channel bandwidth. Doppler fading can be described in terms of the variations in the channel relative to the signal.

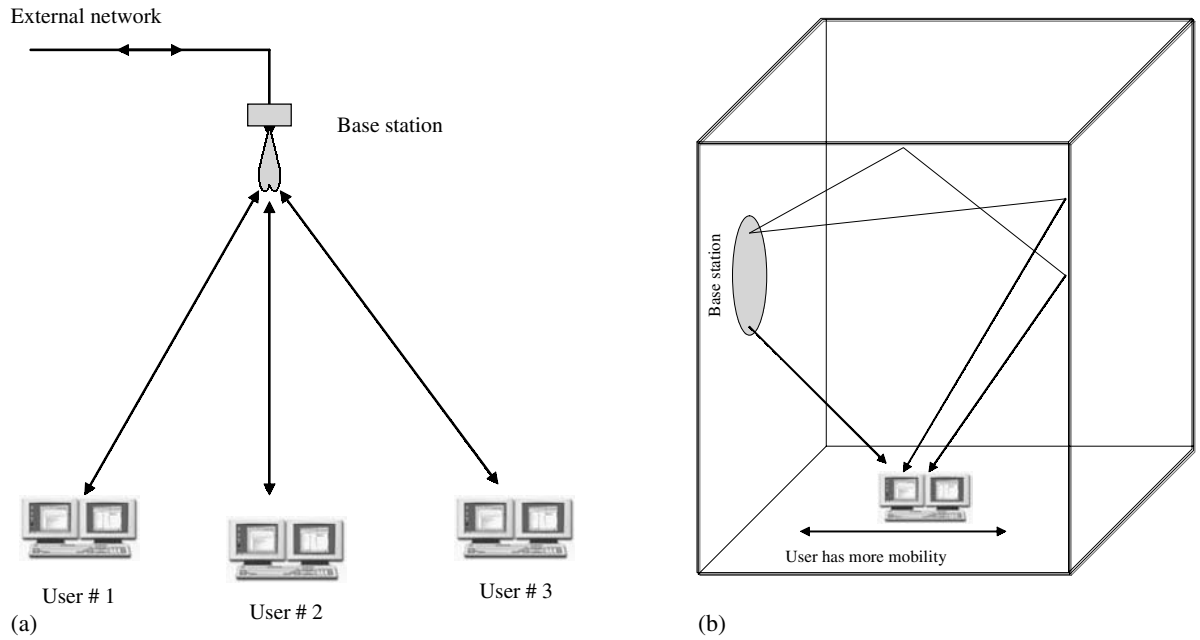


Figure 11: (a) A line-of-sight optical wireless system. (b) An indoor optical wireless system using diffuse light.

create $N\,(N > 1)$ multiple (diverse) independent (at least uncorrelated) versions of the signal. If the probability that signal goes below a threshold is p, the probability that that all these multiple versions of the signals will go below that threshold *simultaneously* is p^N. In other words, if one were to choose the strongest signal from this set, the probability that this chosen signal is above the threshold is $(1 - p^N)$, leading to fading mitigation lessening the chances of outage. The signal processing in terms of choosing the strongest signal is known as selection diversity. Other forms of diversity, such as processing, maximal ratio combining, and equal gain combining, are also available to combine the signals from the diverse branches (Stein, 1987).

Diversity is implemented in the spatial domain (spatial diversity) by having multiple transmitters or multiple receivers or in the frequency domain (frequency diversity) by transmitting the same information over multiple frequency bands, and so on. Other forms of diversity, such as angle diversity, polarization diversity, time diversity, and rake receiver, are also available (Shankar, 2001; Stein, 1987).

Although the techniques described here constitute examples of microscopic diversity to mitigate effects of short-term fading, long-term fading or shadowing can be mitigated through macroscopic diversity techniques. Choosing the best base station that provides the strongest signal to serve a subscriber is a form of macroscopic diversity. It is also possible to combine the signals from different base stations to reduce the effects of long-term fading further.

OPTICAL WIRELESS SYSTEMS

Optical wireless networks offer a simple solution to problems of electromagnetic interference in traditional indoor wireless systems (ElBatt and Izadpanah, 2001; IEEE, 1998; Heatley & Neild, 1999; Pohl, Jungnickel, & von Helmolt, 2001). They are also not subject to any license regulation requirements. They are simple to implement so as to cover a large room or area without any signal leaking through the walls or partitions. This confinement of the signal also offers a form of security from snooping from the adjoining rooms. Because most of the systems depend on LOS, fading problems associated with multipath will be less.

The optical wireless systems fall in two groups, one using LOS systems and the other using diffuse systems. An indoor system based on LOS is shown in Figure 11a.

The signal from the base station reaches the users via direct paths. These LOS conditions exist in the confined space that may also provide a limited mobility and roaming. It is possible to create greater mobility using a diffuse system in which multiple paths exist between the user and the base station. This is shown in Figure 11b. The drawback of the existence of multiple paths is that it leads to the fading condition described earlier in connection with wireless systems in the 900- to 1,800-MHz band.

The indoor optical wireless systems use light emitting diodes and laser diodes operating around 900 nm (900–1,100). The receiver may be a photodiode (p-i-n) or an avalanche photodiode (APD). These systems also have a few major drawbacks. The interference from ambient light is a major problem. The second is the issue of eye safety, which may make it necessary to operate at lower power levels. Because of these factors, the data rate is a few hundred Mbit/s with a range of operation of a few meters (IEEE, 1998).

CONCLUDING REMARKS

This chapter presented brief overview of the topics of interest in wireless systems. Additional reading material may be found in *IEEE Transactions on Wireless*

Communications, IEEE Transactions on Vehicular Technology, IEEE Communications Magazine, and other sources. A number of Web sites containing a wealth of information on some of the topics presented in this chapter are also available at the National Institute of Standards and Technology Web site, http://w3.antd.nist.gov/wctg/manet/wirelesspropagation_bibliog.html (accessed September 5, 2002), and the IEEE Communications Society Web site, http://www.comsoc.org/pubs/pcm/ (accessed September 25, 2002).

APPENDIX: POWER UNITS

The power is normally expressed in decibel units. Power (P_0) in milliwatts (mW) can be expressed in terms of decibel units, dBm, as

$$P_0(dBm) = 10\log_{10}\left[\frac{P_0\,(mW)}{1\,mW}\right]. \tag{A-1}$$

Thus, the power in dBm is an absolute measure of power in mW. 10 mW power is 10 dBm, 1 W is 30 dBm, and 1 μW is –30 dBm. The unit dB, on the other hand, is the ratio of two powers in identical units. For example, if the average signal power is P_0 (mW) and the average noise power is P_n (mW), the signal-to-noise ratio (S/N) can be expressed as

$$(S/N)\,dB = 10\log_{10}\left[\frac{P_0\,(mW)}{P_n\,(mW)}\right]. \tag{A-2}$$

Thus, signal-to-noise ratio expressed in dB carries information on the strength of the signal relative to the noise. If the signal-to-noise ratio is 0 dB, the signal power and noise power are equal. If the signal-to-noise ratio is 20 dB, the signal is 100 times stronger than the noise. If the signal-to-noise ratio is –3 dB, the signal is only 50% of the noise. Loss or attenuation can be expressed in dB units as

$$Loss\,(dB) = Transmit\ power\,(dBm) - Receive\ power\,(dBm). \tag{A-3}$$

If the transmitted power is 10 mW and the received power is 1 μW, one can calculate the transmission loss as $10\log_{10}(10\ mW) - 10\log_{10}(1\,\mu W) = 10 - (-30) = 40\ dB$.

GLOSSARY

Base station (BS) A fixed station in a cellular system that communicates with the mobile units within a cell and may be located at the center (or the edge of the cell). A BS has transmitting and receiving antennas mounted on a tower and is the link between a mobile unit and a mobile switching center.

Bit error rate (BER) The ratio of the number of bits received in error to the total number of bits received.

Cell A geographic region served by a base station.

Doppler shift The upshift or downshift in frequency resulting from the motion of the transmitter with respect to the receiver or vice versa. If the relative speed is v, the Doppler shift is

$$f_d = \frac{v\cos(\theta)}{c}\,f_0,$$

where c is the speed of the electromagnetic wave (3×10^8 m/s), f_0 is the carrier frequency, and θ is the angle between the directions of the transmitter and receiver.

Hand set See mobile unit.

Inter symbol interference (ISI) The interference caused by the overlapping of adjoining symbols or pulses resulting in signal distortion. Viewed in the frequency domain, existence of ISI implies that the channel cannot carry all frequencies with equal gain, with gain decreasing as the frequencies go up.

Link budget The process of computing the maximum transmission distance taking into account the transmitted power, loss or attenuation, and power margin.

Macrocell, microcell, and picocell Microcell typically will cover a few hundred meters while picocells are formed when the base station antenna is located within a building. The coverage of the picocell will be a few meters or less. Macrocells typically cover a few kilometers.

Mobile station (MS) See mobile unit.

Mobile unit (MU) A mobile unit is carried by the subscriber. It may be handheld or vehicle mounted.

Outage Whenever the performance of the wireless system does not reach the minimum acceptable levels, the system goes into outage. For example, if a user requires a minimum power of P_{th} to have an acceptable performance, any time the received power goes below P_{th}, the system goes into outage. The rate at which this happens is the outage probability.

Power margin The excess power budgeted to account for any effects other than the attenuation or the loss of the signal. For example, if a user requires a threshold power of P_{th} (dBm) to maintain acceptable performance and M dB is the power margin to account for fading, the user will require that the minimum acceptable power $P_{\min}$ (dBm) at the receiver be set to $P_{\min} = P_{th} + M$. This has the effect of reducing the available separation between the transmitter and receiver.

CROSS REFERENCES

See *Radio Frequency and Wireless Communications; Wireless Application Protocol (WAP); Wireless Communications Applications.*

REFERENCES

Bertoni, H. L., Honcharenko, W., Maciel, L. R., & Xia, H. H. (1994): UHF propagation prediction for wireless personal communications. *Proceedings of the IEEE, 82,* 1333–1359.

Bultitude, R. J. C., Mahmoud, S. A., & Sullivan, W. A. (1989). A comparison of indoor radio propagation characteristics at 910 MHz and 1.75 GHz. *IEEE Journal on Selected Areas in Communication, 7,* 20–30.

Dersch, U., & Zollinger, E. (1994). Propagation mechanisms in microcell and indoor environments. *IEEE Transactions on Vehicular Technology, 43,* 1058–1066.

Durgin, G., Rappaport, T. S., & Xu, H. (1998). Measurements and models for radio path loss and penetration loss in and around homes and trees at 5.85 GHz. *IEEE Transactions on Communications, 46,* 1484–1496.

ElBatt, T., & Izadpanah, H. (2001). Design aspects of hybrid RF/free space optical wireless networks. In *2001 IEEE Emerging Technologies Symposium on Broadband Communications for the Internet Era* (pp. 157–161). New York: IEEE.

Feher, K. (1995). *Wireless digital communications: Modulation and spread spectrum applications.* New York: Prentice-Hall.

Har, D., Xia, H. H., & Bertoni, H. (1999). Path-loss prediction model for microcells. *IEEE Transactions on Vehicular Technology, 48,* 1453–1461.

Harley, P. (1989). Short distance attenuation measurements at 900 MHz and at 1.8 GHz using low antenna heights for microcells. *IEEE Journal on Selected Areas in Communications, 7,* 5–11.

Hashemi, H. (1993). The indoor radio propagation channel. *Proceedings of the IEEE, 81,* 943–968.

Hata, M. (1980). Empirical formulae for propagation loss in land mobile radio services. *IEEE Transactions on Vehicular Technology, 29,* 317–325.

Heatley, D. J. T., & Neild, I. (1999). Optical wireless—the promise and the reality. *IEE Colloquium on Optical Wireless Communications,* 1/1–1/6. London, UK: IEE.

IEEE (1998). *IEEE Communications Magazine, 36*(12), 70–82.

IEEE (1988). IEEE Vehicular Technology Society Committee on radio propagation, 'Coverage Prediction for Mobile Radio Systems Operating in the 800/900 MHz Frequency Range,' *IEEE Transactions on Vehicular Technology, 37,* 1988, 3–44.

Ikegami, F., Tekeuchi, T., & Yoshida, S. (1991). Theoretical prediction of mean field strength for urban mobile radio. *IEEE Transactions on Antennas and Propagation, 39,* 299–302.

Kennedy, R. S. (1969). *Fading dispersive communication channels*. New York: Wiley.

Jakes, W. C. (Ed.). (1974). *Microwave Mobile Communications.* Piscataway, NJ: IEEE Press.

Lee, W. C. Y. (1986). Elements of cellular mobile radio systems. *IEEE Transactions on Vehicular Technology, 35,* 48–56.

Lott, M., & Forkel, I. (2001). A multi-wall and floor model for indoor radio propagation. *IEEE Conference on Vehicular Technology, 35,* 464–468.

Madfors, M., Wallstedt, K., Magnusson, S., Olofsson, H., Backman, P. O., & Engstrom, S. (1997). High capacity with limited spectrum in cellular systems. *IEEE Communications Magazine, 35*(8), 38–45.

Oda, Y., Tsunkewa, K., & Hata, M. (2000). Advanced LOS path-loss model in microcelleular mobile communications. *IEEE Transactions on Vehicular Technology, 49,* 2121–2125.

Okumura, T., Ohmori, E., & Fukuda, K. (1968). Field strength and variability in VHF and UHF land mobile service. *Review of Electrical Communication Laboratory, 16*(9–10), 825–873.

Pahlavan, K., & Levesque, A. H. (1995). *Wireless information networks*. New York: Wiley.

Pardo, J. R., Cernicharo, J., & Serabyn, E. (2001). Atmospheric transmission at microwaves (ATM): An improved model for millimeter/submillimeter applications. *IEE Transactions on Antennas and Propagation, 49,* 1683–1694.

Parsons, D. (1996). *The mobile radio propagation channel*. West Sussex, UK: Wiley.

Pohl, V., Jungnickel, V., & Von Helmolt, C. (2000). Integrating sphere diffuser for wireless for infra red communication. *IEE Proceedings Optoelectronics, 147,* 281–285.

Rappaport, T. S. (2002). *Wireless communications: Principles and practice* (2nd ed.). Upper Saddle River, NJ: Prentice-Hall.

Rappaport, T. S., & Sandhu, S. (1994). Radio wave propagation for emerging wireless communication systems. *IEEE Antennas and Propagation Magazine, 36,* 14–23.

Saunders, S. R. (1999). *Antennas and propagation for wireless communication systems*. West Sussex, UK: Wiley.

Shankar, P. M. (2001). *Introduction to wireless systems*. New York: Wiley.

Steele, R., & Hanzo, L. (Eds.). (1999). *Mobile radio communications* (2nd ed.). Piscataway, NJ: IEEE Press.

Stein, S. (1987). Fading channel issues in systems engineering. *IEEE Journal of Selected Areas in Communication, 5,* 68–89.

Vogel, W. J., & Hong, U-S. (1988). Measurement and modeling of land mobile satellite propagation at UHF and L-Band. *IEEE Transactions on Antennas and Propagation, 36,* 707–719.

Prototyping

Eric H. Nyberg, *Carnegie Mellon University*

OVERVIEW OF THE SOFTWARE LIFE CYCLE

The conception, implementation, and evolution of software will follow a particular life cycle, depending on the problem to be solved, the software and hardware chosen, and the inevitable changes to system requirements and functionality that happen over time. It can be extremely useful to construct a prototype system in advance of a full-scale implementation; because a prototype typically addresses all of the important aspects of a system to be built, it can be a means to sharpen the system requirements and focus the development team on the most important challenges and risks. In order to understand the role that prototyping can play in an e-commerce application, it is first necessary to understand the basic steps in the software life cycle.

Requirements Elicitation and Specification

The software process begins with requirements elicitation and specification. Through a set of communications with customer representatives, the software engineers begin to specify the system to be built by documenting its desired characteristics: functions provided to the user, user interface, interfaces to other software systems, desired performance criteria, required robustness, error handling and security, update mechanism(s), deployment platform(s), and so on. The result is a formal document, referred to as the Requirements Specification, which is typically used as the basis for detailed planning and contracting with the customer.

System Analysis

During the analysis phase, the engineers model the system by identifying all of the data objects and operations in the problem domain. In object-oriented analysis, domain data objects are represented in an initial class diagram, commonly constructed using the Unified Modeling Language (UML) (Pooley & Stevens, 1999). The functions of the system are broken down into particular interactions with the user, which may imply embedded interactions with remote systems (database lookup, credit transactions, etc.). The system's functions can be represented using a UML use case diagram. For each function or use case, the engineer may write several use case scenarios, prose narratives that describe in detail how the system will carry out a particular function under a variety of circumstances. Use case scenarios include "expected" scenarios, which describe the typical or usual case where no errors or unexpected conditions occur; and "unexpected" scenarios, which cover what might go wrong during the invocation of a particular function (page not found, invalid password, network error, etc.) (Pressman, 2001).

System Design

The design phase maps the functions described during the analysis phase into (a) a set of software modules and (b) a software architecture that integrates the modules into a complete system. At the module level, design has two aspects: (a) partitioning the system's functionality into specific modules (modular design) and (b) specifying the particular algorithms and data structures to be used by each individual module (detailed design). In object-oriented development, modules are typically classes in an object-oriented programming language such as Java. The artifacts produced by the engineer during the design phase might include a detailed UML class diagram, skeletal class files with documentation for data members and methods, and some form of architecture diagram showing how the modules are grouped into partitions (packages) and layers (Bruegge & Dutoit, 2000).

Implementation, Testing, and Maintenance

Once the design is complete, the system developers code the modules according to the design specifications. Implementation should begin with the creation of unit tests for

each module, followed by coding and testing of the module itself. As development progresses, the completed modules are integrated and tested as larger and larger subsets of the system's overall functionality are implemented. The integrated system is tested on the use case scenarios in the requirements specification to ensure that the system behaves as expected in all possible operating environments. As defects are discovered, updates to the system should be carried out using a version-control system to checkpoint stable working versions of the system. Once the system has been delivered to the customer, new releases of the software can be developed and delivered in a similar manner, using previously frozen versions of the system as a baseline.

THE ROLE OF PROTOTYPING IN SOFTWARE DEVELOPMENT

A typical prototype system will involve some level of development in all of the life-cycle activities (requirements, analysis, design, implement, and testing). How much emphasis is given to each life-cycle activity depends on the type of prototype to be constructed and the goals of the prototyping exercise. This section describes the different approaches to prototyping and provides some reasons that prototyping can be a useful first step in system development.

Approaches to Prototyping

The software components created in a prototyping exercise may or may not be part of a full-scale implementation, but a prototype should demonstrate some subset of the desired functionality and user interface. Prototypes can be classified into two broad categories: evolutionary prototypes and throwaway prototypes (Pressman, 2001). Each approach has specific advantages and disadvantages, depending on the goals of the prototyping exercise.

Throwaway Prototyping

A *throwaway prototype* is constructed to demonstrate a simulated capability or user interface, but the software is not intended for use in the final system. Throwaway prototypes can be constructed quickly, using scripting languages (e.g., PERL, Javascript) that may not be suitable for large-scale production. The advantage of a throwaway prototype is that it can be done quickly as a "proof of concept"; such prototypes can be very useful to stimulate discussion with the customer regarding detailed requirements. However, time spent on a throwaway prototype typically does not result in code that can be reused in a full-scale system, so excessive effort spent on a throwaway prototype can be wasteful.

Evolutionary Prototyping

Building an *evolutionary prototype* involves the analysis, design, and implementation of software modules that are intended for use in the final version of the system when it is constructed. An evolutionary prototype might therefore require a significant amount of development (both in time and in programming resources). The advantage of an evolutionary prototype is that "no code is wasted"; it is assumed that most of the effort put into the prototype will result in reusable code that will become part of the final system. However, an early investment in reusable code might not be justified, if the requirements or technology are not yet well understood.

Prototyping PROs and CONs

In an effort to streamline development schedules and minimize development costs, a project manager may be tempted to re-use code built during a throwaway prototyping exercise. This can wreak havoc on the development of a full-scale system, because a throwaway prototype may have been coded quickly, with no attention to proper analysis, design, or careful implementation. The platform used for a throwaway prototype may be completely inappropriate for a full-scale production system. For example, a throwaway prototype might be constructed using PERL scripts, flat data files, and hand-coded HTML, where a production-quality system might require the use of a session architecture, a relational database, and dynamically generated Web pages. Another characteristic of prototypes is that they almost always leave out certain functions or capabilities of the full-scale system. A prototype design that encompasses only a subset of the desired functionality may not be easy to extend to a full implementation. For example, a prototype which delivers Web content in a single language might not utilize a design that supports easy extension for multilingual content; if the full-scale system requires localization for several countries and languages, it will be necessary to re-design the data architecture for a full-scale system. In general, if the requirements aren't well understood at the outset, or the technology is new for the development team, then a throwaway prototype is appropriate, but management must be sensitive to the risks associated with reuse.

On the other hand, it may be tempting to declare that all prototyping will be of the evolutionary variety, so that "no effort will be wasted." This is a mistake if the problem or the technology isn't well understood, because the analysis and design may incorporate defects or misunderstandings that should be resolved before a full-scale system is built. When the team is experienced with the problem domain and the software technology, this risk is reduced, but it can be difficult to guarantee future reuse before construction and test of an initial system. When a new category of system is constructed for the fist time, prototyping becomes part of a "generate and test" paradigm for system design; design ideas are formulated and initial systems constructed as a means of evaluating tentative design decisions. In general, when the requirements aren't well documented, it is usually more appropriate to construct a throwaway prototype that can serve as the basis for a customer review and discussion regarding the emerging requirements and system design.

Why Do We Do Prototyping?

There are a variety of reasons to construct a prototype; most involve a desire to better understand the problem domain, the user's requirements, the technology to be used, or the trade-offs inherent in various system design choices (technology, architecture, etc.). Prototyping

is also extremely important in the presence of various software risks that may not be well understood.

Requirements Analysis and "Design by Example"

Requirements analysis for e-commerce systems presents two main challenges for the engineer: (a) it is generally difficult for a customer to provide a detailed specification for the visual components of a Web site (graphics content, layout, navigation, etc.); and (b) there are embedded functional layers which the customer may take for granted (database access, credit transactions, order fulfillment, etc.). Once a prototype has been constructed, it is fairly easy for the customer to provide feedback on what he or she likes or does not like about the proposed design of the interface. The construction of a prototype can also uncover a variety of assumptions regarding integration with existing systems; in reviewing a prototype Web site, the customer may realize that certain legacy functionality is missing, links to other corporate sites are required, other business units must be represented, etc. The customer is typically less well informed about the technical requirements of an application; the prototyping exercise can be used as a vehicle for initiating contact with all of the people "behind the scenes" who understand the low-level integration requirements of a particular application, including details of existing CRM and ERP systems, security requirements, etc.

Understanding the Problem Domain

The software development life cycle proceeds smoothly when the problem domain is well understood. When there are gaps in the engineer's understanding, even the best practice can fail to produce the desired result. When a development team begins working in a new area, prototyping is a useful way to immerse the team in the particulars of the problem before the team members are asked to design and implement a full-scale system. For example, a development team that shifts from the creation of static Web sites to the creation of shopping portals can take advantage of a prototyping exercise to learn the functional requirements of a shopping portal in detail. An analogous situation arises when a development team must learn a new technology; even with a relatively stable technology (e.g. Java), it is useful for the team to undertake a prototyping exercise as preparation for a full-scale project.

Reducing Technical Risk

Technical risks arise when new technology is used for the first time, or a novel combination of technologies is tried for the first time. In either situation, there is a significant probability that unexpected problems will arise, and a prototyping exercise can uncover these unknowns and improve the accuracy of time and cost estimates for a full-scale system. In the worst case, an initial technology decision may be abandoned as the result of a failed prototype, but this is clearly preferable to canceling or renegotiating a large-scale development contract that is already under way. In the world of e-commerce software development, technical risk is one of the biggest unknowns. As new paradigms are being created for distributed client–server solutions over the Internet, companies are compelled to adopt new technology before it has been widely tested, in order to remain competitive. Knowing when to invest in rapid prototyping to alleviate technical risk is an important strategic skill.

THE NATURE OF E-COMMERCE SOFTWARE

Software development is an inherently risky endeavor. Complex systems are difficult to deliver without significant defects, even when adequate time and resources are available. E-commerce systems are even more challenging, because they are subject to additional pressures that increase the risks associated with software delivery.

Development on "Internet Time"

Unlike company-internal software development, which proceeds according to schedules set within the organization, e-commerce systems are often developed under the pressures of "Internet time"—the expectation is that new technologies and solutions are rolled out every 3 to 6 months, under the credo "evolve rapidly, or die." This time pressure has an adverse effect on software development. In the absence of adequate time to complete the various phases in the software engineering life cycle, software quality degrades as less time is spent on analysis, design, and testing. With so little time available for development, one might question whether prototyping is an appropriate strategy, especially if a throwaway prototype (and its associated "waste" of resources) is being considered. On the other hand, one might argue that prototyping is absolutely essential to keep the development team on top of new technology, so that full-scale applications can be designed, implemented, and deployed as quickly as possible once the requirements are stabilized (Iansiti & MacCormack, 1997). Because prototyping can alleviate technical risk, it can improve the responsiveness of the development team and help to keep final product delivery on schedule, even if all requirements are not known until close to the planned delivery date.

Evolution and Obsolescence

When software is developed quickly with an ad hoc or nonexistent software process, the result is often code that is poorly designed, poorly documented, and difficult to maintain. When faced with a new project, the development team is likely to resist working with the old system, and will prefer to write a new system from scratch. The pressures of rapid development can feed this vicious cycle, and the result is a significant amount of time wasted on reimplementing modules when previous modules could have been reused—assuming they had been well-designed and well-documented in the first place. If a prototype is constructed in advance, a code review can be utilized to identify which portions of the system are likely to be general (and hence reusable). When the full-scale system is constructed, the design and implementation of those modules can be done in a manner that maximizes their future reusability. In the future, the most successful software companies will be those that minimize the amount of new code that is created in each new application, through

careful design and reuse of generic code libraries. This is difficult to achieve if every application the company constructs is a "first system."

> Your plans and your technology must be as fluid as the changing environment in which they operate. Your software, your systems, your entire technology architecture must naturally mold and adapt to changes without costly, time-consuming infrastructure overhauls. Your e-business software platform must provide a safe, yet powerful catalyst for ongoing innovation, increasing customer value-add, and continuing competitive advantage. (BEA Systems, 2002)

Architectural Complexity

A great number of e-commerce systems adopt the multitier approach to software architecture; the simplest (and most common) variation integrates a Web browser, a Web/application server, and a back-end database. If we think of these subsystems as residing in different architectural layers, then the boundary between each layer is an appropriate place for security measures such as firewalls and virtual private network routing (Zwicky, Cooper, & Chapman, 2000). When an e-commerce architecture is constructed for the first time, an initial prototype can be used to understand the complexities of integrating the various layers and their associated security protocols. A prototype is also a useful vehicle for analyzing the properties of a proposed architecture, such as

Scalability. How feasible is it to add multiple application servers or mirror the backend database? The proposed design can be evaluated in the context of the prototyping exercise, by attempting to show how multiple servers could be added to the system for load balancing, mirroring, etc.

Robustness. How will the system react under high network load and/or slow response times? What will happen if a particular networked service goes down? Once a prototype system has been constructed, the team can build test routines that simulate various operating conditions. It is useful to benchmark a Web application server by simulating an increasing number of simultaneous user accesses in order to understand how average response time degrades under increasing load. Simulation can also be used to test how the system responds when certain services become unavailable to overload conditions (network timeout) or system crashes.

Security. What points in the architecture are vulnerable to attack? Can we determine them using attack simulation software? If a prototype is constructed in a realistic operating environment (including firewalls, etc.) then simulation can be used to identify and remedy potential system vulnerabilities.

Integrity. Are there any circumstances where the user's data may be lost? To answer this question requires a solid understanding of scalability, robustness, and security characteristics of the system and how they impact the functional requirements. For each functional requirement or supported user action (login, place order, etc.) the prototype can be analyzed to determine whether known issues with robustness and security might impact the functional requirement. Information gathered from prototype evaluation can be extremely valuable in refining a system design before a full-scale system is constructed.

Technological Risk

During product development, an organization must often select between competing technologies (such as programming languages or program libraries) to implement parts of the product architecture or functionality. For example, in early 1996 NetDynamics had the choice of adopting the Java programming language or writing its own proprietary language for a major release of a database integration product (Iansiti & MacCormack, 1997). At that time Java showed promise as an emerging standard, but was still an immature platform for large-scale development. The engineers at NetDynamics developed a series of prototypes in order to understand the benefits and risks associated with each choice. The development drawbacks of working with an immature language (which may lack a stable, robust development tool suite) must be balanced against the longer-term advantages for the application programmer. This trade-off exists with every emerging technology, and iterative prototyping provides an effective means of evaluating the associated risks in each case.

PROTOTYPING FOR E-COMMERCE SYSTEMS

The preceding sections provide a motivation for prototyping in e-commerce system development. This section presents the various aspects of the e-commerce software life cycle that can benefit from prototyping, followed by specific life-cycle models that are adapted for Web-based applications and flexible product development on "Internet time."

Use Case Scenarios

A modern e-commerce system brings together content and functionality on several levels. In order to understand the overall flow of information in the system and how the various system modules must interact, it is essential for the engineer to analyze the different patterns of usage the system is expected to support. A use case analysis will identify the different users of the system (customer, installer, maintainer, content provider, administrator, etc.) as well as the functions to be provided for each user. An initial prototype should include a skeletal implementation or simulation of the interfaces and functionality for each user role. Following the construction of the prototype, a requirements review can be conducted with prospective members of each user group to determine whether the current understanding of the requirements (as embodied in the prototype) is complete and correct. At the earliest stages in development of an e-commerce site, it is useful to build a prototype of the system's front end (Web site,

applet, or application) as a tool for reaching consensus on the system's modes of operation, inputs, and outputs.

Object-oriented analysis (Bruegge & Dutoit, 2000; Coad & Mayfield, 1999) and the Unified Modeling Language (Pooley & Stevens, 1999) provide useful tools for use case analysis and requirements modeling. However, the resulting model diagrams cannot be discussed directly with a customer or user (unless he or she is skilled in UML, which is unlikely). It is more useful to construct a prototype that illustrates an initial version of each user function in a familiar interface (e.g., Web pages or applets). Once such a prototype has been constructed, the engineers can demonstrate the system for the customer, talking him or her through each proposed function for each user role. Feedback from this kind of prototype demonstration can be revealing. Often the initial requirements from the customer are vague or ambiguous; a prototype review can uncover potential misunderstandings. For example, an initial requirement might be specified as "provide a product search capability for the user," without being specific about the search criteria. If the prototype omits certain search criteria (e.g., particular product characteristics that are important to the customer), the customer will notice in the review, and the requirements document can be updated to reflect the more specific requirement.

Interface Prototypes

A typical Web-based system will include several pages or "screens" in support of each user interaction or user function. Each of these pages will contain a variety of user interface elements, such as hyperlinks, pull-down menus, fill-in forms, checkboxes, and radio buttons. The pages will be arranged in a particular hyperlinked structure, starting from the home page of the application. If possible, navigation should be transparent and effortless, so that the user can focus on the task at hand without being distracted by poorly designed navigation, or content that is difficult to scan visually (Nielsen, 2000). A prototype can be used to evaluate a set of criteria regarding the interface; questions to consider include the following:

Are the user interface elements the most appropriate ones for the task? In order to minimize the time it takes for the user to scan a Web page and make the appropriate selection or provide required data, it may be necessary to redesign the layout and interface components. For example, a long block of text with choice buttons at the bottom will be less effective than a short list of bullet items including hyperlinks. The former must be read in its entirety, whereas the latter can be scanned quickly for the desired choice. Fill-in forms should be constructed for easy visual scanning from field to field, and common conventions (e.g., tab key) should be utilized for switching the focus from field to field.

Do the user interface elements support an effective task flow? If a user task involves decision-making during a set of sequential steps, it is important not to force an early decision that might be retracted later based on the result of a subsequent step. Prototype evaluation should consider what will happen if the user changes his or her mind about the task at hand. Normal progress through a task should be streamlined and effective, but it must also be easy for the user to retract a choice, cancel a transaction, or start over.

Are suitable defaults provided for menu selections and fill-in forms? An initial requirements specification may not be detailed enough to provide information about the default content or behavior of fill-in fields, search parameters, date ranges, etc. which are part of a user input to the system. Effective defaults are an important part of streamlining the user's experience, especially when a lot of data must be entered and sensible defaults exist.

Does the system check for combinations of data values that are invalid for the application? It should not be possible for the user to enter combinations of values in a fill-in form that result in an invalid request or transaction. Testing a prototype can be an effective way to uncover such cases, if they have not been specified in detail in the initial requirements documentation.

Is it easy to navigate from page to page in the site or application? Effective Web applications are "training free"—the user can immediately intuit how to navigate through the site or application after briefly scanning the first page or screen. If the user has to stop and think about how to navigate, or worse, experiment with different mouse clicks to find what he or she needs, the effectiveness of the site will suffer and the user's subjective experience will degrade.

Can each part of the site or application be reached with a minimum number of mouse clicks? Even if intuitive navigation is provided, a particularly deep or content-rich site might require lots of user interaction to reach certain portions of the site content. A prototype evaluation can assess whether a reorganization of the site content or navigation structure might provide better accessibility. Secondary navigation mechanisms such as search engines or site maps can be valuable in this regard.

Although by no means exhaustive, this list illustrates how a prototype review can be used to improve the interface component of a Web application. By discussing these criteria with the customer and/or evaluating them in a more formal usability study, valuable information can be gathered which can trigger design improvements for a full-scale system.

Content Prototypes

The term *content* is used in an e-commerce context to refer to the parts of the site or application that are not directly related to navigation or data entry, which are typically presented as a combination of text, graphics, and/or multimedia. Content must simultaneously satisfy two basic criteria—it should be consistently presented and it must be easy for the user to scan and understand. The page designer determines how to partition and present information in a manner that is consistent with the overall visual design (look and feel) of the site. Choices of typeface, font size, spacing, color, and page size all have a dramatic effect on how easy it is for the user to assimilate the content (Siegel, 1996). Helping the user to complete the task

at hand is typically more important than sophisticated graphic design. The text should be optimized for scanning, not reading; short text segments, hyperlinks, and bulleted lists will allow the user to scan the current page quickly and determine his or her next action (Nielsen, 2000). The content must also meet the goals of the application. If the system is designed to support online sales of a particular product line, then the relevant content for each product should be available. Although prototype systems typically focus on determining the look and feel of a site via graphic design, it is clear that content should also be reviewed to ensure that it is optimized for usability.

In cases where multilingual or multicultural content is required, an additional set of criteria must be considered:

Translation of content. If the site or application is intended for global access, then it might be necessary to provide site content in more than one language. If this requirement is adopted, then for each language to be supported the system design must consider (a) which version or dialect to support, (b) which character sets to use, (c) which font encoding to use, and (d) how to provide text, graphics, and HTML in that language (O'Connor, 2001). Most translation work is handled by selection of an appropriate service bureau. It is useful to construct a multilingual prototype system, as this affords an opportunity to evaluate the quality of service provided by potential translations vendors. A prototype can also be evaluated for how well the site or application presents multilingual content to the end user; the selection of a language at run-time should be easy (or automatic). Sites such as Google (http://www.google.com) predict the desired language from the IP address of the Web browser when it contacts the site and redirect the HTTP request to an appropriate language server. Emerging standards such as Unicode provide a standardized way of encoding Web pages for multiple languages and character sets (Unicode Consortium, 2002).

Localization of content. Preparing content for a different language often implies preparing it for a different culture. *Localization* is the process that adjusts the translated material to fit the targeted culture. Simple forms of localization include changing the page layout to fit the translated material; because translated material can vary in length, the original layout may no longer be appropriate. More sophisticated forms of localization involve checking for culturally inappropriate use of graphics or other visual cues (for example, the use of a black border is typically reserved for death notices in the Japanese culture; use of a black border might be inappropriate in most other contexts). It is also possible that localized versions of a site, application, or service might differ in basic content or scope of service; not all information and/or functionalities may be available in all markets. Building and evaluating a prototype system is an excellent way to determine the exact nature of the localization requirements for a particular global market.

Infrastructure support. Typical e-commerce Web sites include various navigation mechanisms, application servers, back-end databases, external servers and services, etc. All of these layers in the computing infrastructure must be considered in designing a site for the global market. If a decision is made to translate and localize content and services, then the corresponding components in the infrastructure (product database, order fulfillment, etc.) must also be translated and/or localized. A Web site that provides product information in several languages but provides only a single language for check-out is not responsive to its customer base. It is much better to catch this type of oversight in a prototyping evaluation than to have it discovered by customers online.

Content maintenance. Although it is a challenge to launch a multilingual site, the ongoing cost of maintaining a site in multiple languages can also be significant. It is useful to identify, prototype, and evaluate a process for content update before a site "goes live." Ideally, translated content will be available as soon as primary content has been updated, but this can be difficult to achieve if the organization relies on outsourcing for translation and localization. Although automatic translation tools can be useful as human aids, the state of the art is still not quite good enough to be relied upon without human intervention, except in very narrow information domains. In addition, the tools used for page creation and overall site maintenance must be able to support multilingual content. Software that keeps track of the relationships between primary content pages and their translated equivalents is a must.

As it is typically very costly to create and provide content in multiple languages, it is important to ensure that the design of the site or application is sound before it is scaled up to include all target languages and markets. Prototyping can be an excellent way of creating and evaluating a design for multilingual content delivery, before full-scale development begins.

Architecture Prototypes

E-commerce architectures typically involve multitier integration of browsers, applets, servlets, application servers, and back-end services (databases, credit transactions, etc.). Most of the complexity (and most of the difficulty) in e-commerce development is associated with the integration and testing of the various components in the architecture. Even before the full content for the application has been prepared, it is useful to build a prototype implementation of the end-to-end system. As mentioned above, architectural prototyping can serve as the basis for a variety of tests, including the following:

Integration Testing. Do all of the components communicate as expected? A Web application typically relies on different Web browsers communicating with a Web server and/or application server; these in turn communicate with various back-end services (product database, order database, etc.). It is important to test assumptions regarding the connectivity of the chosen software and hardware for the project, as early as

possible. A prototype system with limited functionality (just a few services and limited content) can be used to test the integration of the chosen technologies.

Performance Testing. What is the average time required to service a typical user transaction? Is the system's response time acceptable? How does system performance vary as the transaction load is increased (e.g., in a simulated test environment)? Once a prototype has been constructed, the architecture can be tested via simulation of an ever-increasing number of page and transaction requests. In the course of such an evaluation, resource bottlenecks (such as an overloaded server) can be identified and addressed. If necessary, the original design can be revised to better support scalability (redundant servers with load balancing, mirrored databases, etc.).

Security Testing. Is the system vulnerable to any known exploits that could compromise system performance or data integrity? Attack simulation software can be used to test the security of the site once it has been prototyped. It is far better to uncover and address security problems before the full-scale site has been built and launched.

Early prototyping can also be used for comparative evaluation of competing technologies and to evaluate whether specific emerging technologies are appropriate for a given application. Several commercial Web application environments are available, and each makes certain assumptions about the operating environment (operating system, platform, etc.). How these application paradigms fit within a particular organization and integrate with existing legacy components is an important area for early investigation.

A Life-Cycle Model for Web Projects

An effective development model for Web-based e-commerce systems must support rapid development and continuous evolution. Web applications in particular are subject to constant improvement and content update. Copywriters, graphic designers, and Web designers provide content. Software engineers and developers provide the technical infrastructure. It is common that both content and technology are developed, deployed, and maintained simultaneously and somewhat independently. Pressman (2001) proposes the WebE process, an iterative life-cycle model that acknowledges the need to develop content and technology in parallel. The WebE process includes the following cyclic activities:

Formulation. Identify the goals of the application, the users, and the scope of both content and function. This process is like the normal requirements elicitation and specification in typical software engineering, but with important differences. It is important to identify all stakeholders associated with the application, as well as the content providers. A typical application requires assimilation of content from different organizational units, from print and online resources, in varying formats. Identifying these requirements early is important for an accurate estimate of the effort to build and maintain the site, which could involve significant amounts of content conversion.

Planning. Allocate appropriate resources and decompose development into appropriate phases, milestones, and schedules. This process must take into account special requirements regarding content acquisition, conversion, translation, and localization.

Analysis. Understand the details of the system's content, user interactions, specific functionalities, and software configuration. For each of these areas, construct an analysis model that identifies all of the data objects and operations on those objects.

Engineering. Design and implement the individual elements of the system. Content creation and system implementation take place in parallel. Content creation begins with content design, followed by a content production process that encompasses acquisition, conversion, translation and localization. System implementation includes architectural design, navigation design, and interface design.

Page Generation and Testing. Construct the actual pages for the site. All of the individual components (HTML and JSP pages, applets, servlets, CGI scripts, etc.) are integrated and tested by following all of the steps represented in the use case scenarios built during requirements formulation.

Customer Evaluation. Perform a detailed review of the application with intended users and gather feedback for next version of system.

Although the full set of WebE activities is appropriate for development of a complete application, the same steps can also be followed in the construction of an initial prototype system, with selection of an appropriate (reduced) scope. In projects where I have used this approach for prototyping, the initial focus is on the analysis and engineering steps. The team creates a set of analysis models (use case diagrams, usage scenarios, and a class diagram) and embodies them in a set of prototype Web pages that capture a proposed content style, navigation method(s), and user interface elements. The architecture is usually minimized or simulated. For example, the functions of the application server and product database may be simulated by simple CGI scripts. The customer evaluation step focuses on whether the proposed page design and navigation capture the desired task flow and information objectives. Customer feedback helps to refine the analysis models before development begins.

A second iteration typically involves the creation of an architecture prototype, where the prototype content and navigation are deployed in an end-to-end system that includes the final application server and database modules. The architecture prototype must contain only enough information content to test end-to-end integration; scaling the system up to full content can happen in subsequent iterations.

The WebE Process is flexible enough to support any number of successive prototypes or incremental versions leading up to a final product, while reinforcing the notion that each revision to a system should take place in a software engineering context. Each evolution includes

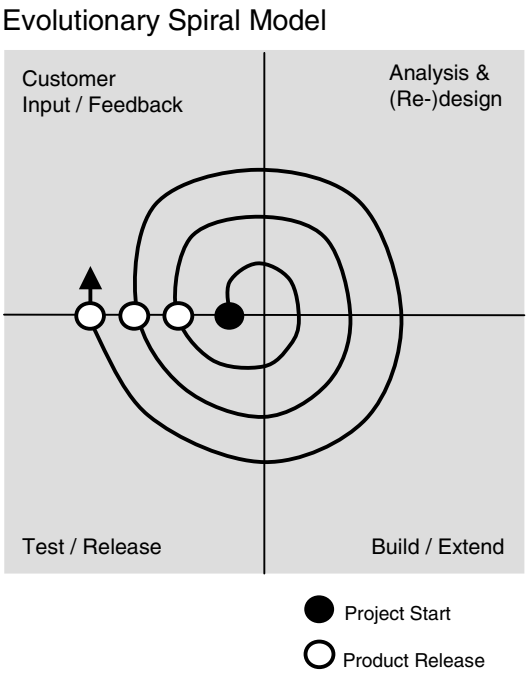


Figure 1: The spiral development model.

updates to the system's specifications, evaluation of existing modules for reuse, careful engineering and testing, and subsequent customer evaluation. Even if initial prototypes are not extended or reused, analysis and design can produce specifications that are useful for a full-scale implementation. Iterative prototyping fits a general form of the Spiral Model, a development process model first proposed by Boehm (1988); see Figure 1.

Flexible Product Development and the Evolutionary Development Model

Web-based applications are just one example of systems that are developed rapidly and evolve constantly over time. Many other types of applications are created in dynamic, unstable environments where technology and customer requirements are in constant flux. High-quality software can be developed in such an environment with the use of iterative, evolutionary prototyping in a flexible development process.

Early Prototyping in Product Development

A study of 29 completed development projects (MacCormack, Verganti, & Iansiti, 2001) showed that the creation of an early, incomplete prototype for customer evaluation is an important predictor of software quality in dynamic, unstable environments. After release of the first prototype, the development team must work to integrate the customer's feedback into the software design as it evolves towards the final deliverable. This implies an evolutionary development model, where active development of the final product begins before the final requirements are known. A series of evolutionary prototypes may be constructed and evaluated before the requirements are frozen for a major release of the product. Evolutionary development will be more successful when the team has made a substantial investment in a flexible, modular architecture that supports rapid adjustments to an evolving product design. Team members with past experience on several different projects and software versions are more likely to thrive in an evolutionary development environment.

A Flexible Development Process

The total time from project start to market introduction (*lead time*) can be divided into *concept time* and *response time* (Iansiti & MacCormack, 1997). Concept time is the window of opportunity for introducing requirements into the system concept, and matching the technology with users and application context. Response time is the time taken to produce the deliverable system after the system concept is frozen. In a traditional software development process, system implementation does not begin until all the requirements have been frozen. In a more flexible model, implementation (e.g., evolutionary prototyping) begins before the requirements are frozen, and response time can be shortened significantly. When several incremental product versions are to be developed, it is important to minimize the response time (during which requirements are frozen). In the most dynamic formulation, a flexible development process will constantly acquire new information about customer requirements, test alternative design solutions, and integrate both into coherent, high quality products. In a study by MacCormack (2001), the software system with the earliest beta release and the greatest number of beta releases—Netscape Navigator 3.0—also received the highest score for perceived software quality. See Figure 2.

Infrastructure Requirements

The flexible development process requires an investment in development infrastructure, primarily in version control and testing. Organizations that have demonstrated success with rapid product cycles, such as Netscape and

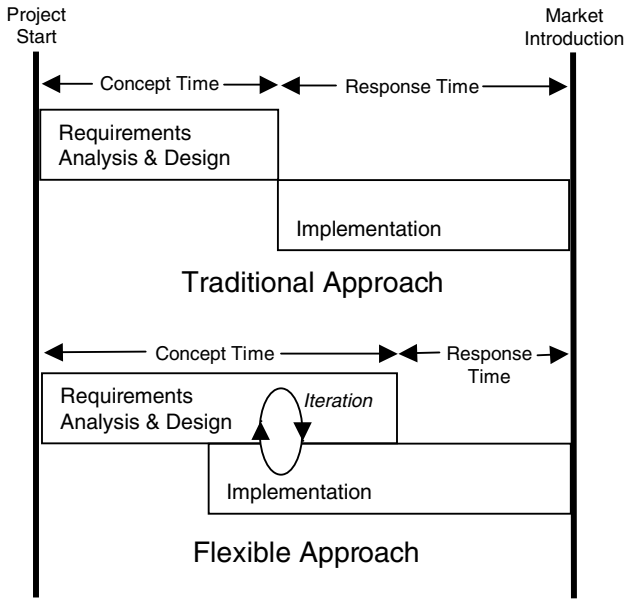


Figure 2: A flexible approach to development.

Microsoft (MacCormack, 1991), decompose a product into several subcomponents or "microprojects" that contribute to the overall functionality. Delivery of an early prototype requires that critical functionalities can be developed early, before all product functions are in place. This presupposes an ability to decompose the system into functional modules, which can be replaced by dummy code or simulated behavior before full implementations are available. In order to manage the complexity introduced by parallel development of different modules and different product releases, strong version control is necessary. Effective version control involves the use of a version control system (e.g., CVS, 2002) to manage and track updates to the source code, along with a test process that promotes global coherence of the overall product during parallel development. The *synchronize and stabilize* model developed by Microsoft (Cusamano & Selby, 1997) requires frequent synchronization and testing of the entire system under development. Intermediate testing is most effective when developers receive feedback as quickly as possible after a system update (MacCormack, 1991).

CONCLUSION

Prototyping is an important part of software engineering. Building a prototype can be an effective way to build customer consensus on requirements, gain clarity on the details of a complex architecture, determine the look and feel of a Web site, and alleviate the technical risk associated with emerging technologies. Although e-commerce developers are under pressure to produce working systems quickly, the use of an incremental prototyping model helps to improve the quality of delivered systems by uncovering defects and performance problems earlier in the development cycle. An iterative Web life-cycle model (such as Pressman's WebE process) involves all stakeholders early in the development process and gives explicit emphasis to content design and creation, which are a distinguishing (and challenging) characteristic of Web applications. A continuous prototyping effort can begin with content and navigation prototyping, followed by prototyping on an end-to-end architecture. At each step of the way, the existing requirements and design can be evaluated and updated to better reflect the emerging consensus about what the system should do, and how it should do it. Flexible development models, evolutionary prototyping, and component-based software development are important trends for ongoing and future development of e-commerce systems.

GLOSSARY

Analysis An activity in the software life-cycle that identifies and specifies the data objects and operations in a particular problem domain.

CRM (customer relationship management) The business process that captures and stores information about the customer, including preferences, past and present orders, business relationships, etc.

Content The information conveyed by a Web site, such as company information, product information, news releases, order history, and user preferences.

Design An activity in the software life-cycle that identifies and specifies a set of software modules that will be used to implement the data objects and operations in a given problem domain.

ERP (enterprise resource planning) The global management of business processes such as procurement, administration, project management, sales force automation, product development, and order fulfillment.

Evolutionary prototyping A prototyping effort that produces a software system where some (or all) of the modules can be reused as part of a full-scale, deployable application.

Java An object-oriented programming language that is often used to implement e-commerce software systems.

Life cycle The entire set of activities surrounding the conception, design, creation, testing, deployment, and long-term maintenance of a software system.

Problem domain A term used to describe the elements of the software operating environment, including users, maintainers, legacy systems, and interactions between users and the system.

Prototyping The creation of an initial version of a software system, often with limited functionality, to support problem analysis, requirements definition, interface design, etc.

Requirements The formal elements of a problem definition that must be satisfied by a software system; these can include functional capabilities, performance capabilities, platform capabilities, integration capabilities, etc.

Software architecture A description of how a set of software modules will be deployed, how they interact with each other, how they will be installed on various hardware devices, and how they will interact with existing systems and users.

Software complexity A measure of how difficult a software system is to design, implement, and deploy; systems that have a greater number of functional requirements, interactions with users, communications with legacy systems, etc. tend to be more complex.

Software evolution The process of maintaining a software system from initial deployment through retirement from active use; evolution involves updates to support operating system changes, changing user requirements, performance criteria, etc.

Software obsolescence Reached when a software system no longer meets the changing requirements of its users and can no longer be extended or maintained in a cost-effective manner.

Technical risk The uncertainty inherent in the use of new technology that has not been widely tested or in the novel use of existing technology in a new problem domain.

Throwaway prototyping A prototyping effort that produces a software system that should not be reused as part of a deployable application.

UML (unified modeling language) A graphical notation for analysis and design models that unambiguously specify the requirements and design elements for a software system.

Use case scenarios Textual descriptions of the functional requirements that must be met by the software; typically, scenarios outline all of the interactions with the user and other external systems.

User interface That portion of the software system which is visible to the end user; for e-commerce systems, this is often a Web page or applet viewed in a Web browser.

CROSS REFERENCES

See *E-business ROI Simulations; Java; Return on Investment Analysis for E-business Projects; Risk Management in Internet-Based Software Projects; Software Design and Implementation in the Web Environment*.

REFERENCES

BEA Systems, Inc. (2002). Future-proofing your business. Retrieved June 25, 2002 from http://www.bea.com/products/ecommerce_wp.shtml

Boehm, B. (1988). A spiral model of software development and enhancement. *IEEE Computer 21*, 61–72.

Bruegge, B., & Dutoit, A. (2000). *Object-oriented software engineering*. Upper Saddle River, NY: Prentice–Hall.

Coad, P., & Mayfield, M. (1999). *Java design*. Upper Saddle River, NY: Prentice–Hall.

Cusamano, P., & Selby, R. (1997). How Microsoft builds software. *Communications of the ACM, 40*(6), 53–61.

CVS (2002). *Concurrent versions system: The open standard for version control*. Retrieved November 19, 2002 from http://www.cvshome.org

Iansiti, M., & MacCormack, A. (1997, September–October). Developing products on Internet time. *Harvard Business Review, 75*, 108–117.

MacCormack, A. (2001). Product-development practices that work: How Internet companies build software. *MIT Sloan Management Review, 42*(2), 75–84.

MacCormack, A., Verganti, R., & Iansiti, M. (2001). Developing products on Internet time: The anatomy of a flexible development process. *Management Science, 47*(1), 133–150.

Nielsen, J. (2000). *Designing Web usability*. Indianapolis: New Riders Publishing.

O'Connor, J. (2001). Creating a multilingual Web site with Apache. *Multilingual Computing, 12*(15), 56–58.

Pooley, R., & Stevens, P. (1999). *Using UML*. New York: Addison–Wesley.

Pressman, R. (2001). *Software engineering: A practitioner's approach* (5th ed.). New York: McGraw–Hill.

Siegel, D. (1996). *Creating killer Web sites*. Indianapolis: Hayden Books.

Unicode Consortium (2002). *What is unicode?* Retrieved August 22, 2002 from http://www.unicode.org/unicode/standard/WhatIsUnicode.html

Zwicky, E. D., Cooper, S., & Chapman, D. B. (2000). *Building Internet Firewalls*. Cambridge, MA: O'Reilly Associates.

FURTHER READING

Nielsen, J., & Tahir, M. (2002). *Homepage usability: 50 Websites deconstructed*. Indianapolis: New Riders.

Nyberg, E. (2001). *Multilingual Web sites: Trends and technologies*. Executive Education Presentation, Institute for E-Commerce, Carnegie Mellon University. Retrieved August 22, 2002, from http://www.cs.cmu.edu/~ehn/ExecEd

Sommerville, I. (2001). *Software engineering*. New York: Addison–Wesley.

Van Duyne, D., Landay, J., & Hong, J. (2003). *The design of sites: Patterns, principles and processes for crafting a customer-centered Web experience*. New York: Addison–Wesley.

Public Accounting Firms

C. Janie Chang, *San José State University*
Annette Nellen, *San José State University*

INTRODUCTION

Internet and E-business Overview

In the last decade, the Internet and World Wide Web (WWW) have changed the way people communicate, conduct business, and manage their daily lives. Early development of the Internet was supported by the Defense Advanced Research Project Agency (DARPA) of the Department of Defense (Greenstein & Vasarhelyi, 2002; Kogan, Sudit, & Vasarhelyi, 1998). The Internet came online in 1969 for scientists and researchers to keep in touch with one another and to share each other's computers (Deitel, Deitel, & Steinbuhler, 2001). At that time, the major benefit of the Internet was the capability of quick and easy communication via e-mail. The introduction of the WWW occurred in the early 1990s. Part of the Internet, the WWW allows users to locate and view multimedia-based documents (Awad, 2002).

Advances in hardware and software have led to explosive growth of the Internet and the WWW. In addition to launching us into the information age, electronic commerce (e-commerce) brings universal access via the Internet to the core business processes of buying and selling goods and services. The Internet's worldwide reach helps businesses discover new markets and increases the speed of access. Electronic business (e-business) is a broader term than e-commerce. E-business includes not only the exchange of information related to the actual buying and selling of goods or services over the Internet but also the circulation of information among employees and other external information users (Greenstein & Vasarhelyi, 2002).

Given the nature of a certified public accounting (CPA) firm's activities—tracking thousands of pieces of data, the need for up-to-date information, trusted client relationships, and high-quality service—broader usage of technology has long provided a benefit to CPAs. Various types of technology have led to increased efficiencies in CPA firms and have made it possible for the firms to provide new and enhanced services to their clients. The Internet and e-business applications will continue to improve and expand the range of services offered by CPA firms.

Technology Trends for the Accounting Profession

Every aspect of the accounting profession is being pervasively affected by advances in information technology (IT). IT has made CPA firms efficient in performing most accounting, audit, and tax functions more accurately, reducing the amount of paper they must maintain, and expanding the range of services they offer to clients. Time saved by using IT in making rote calculations (such as calculating depreciation expense), calculating projections, preparing tax returns and financial statements, and performing audits enables CPA firms to have more time to provide consulting services and personal attention to clients. Via the Internet, CPAs have increased mobility; that is, they have access to client information and reference materials 24 hours a day, 7 days a week (24/7), and from any location, which allows them to better serve clients and to have clients in many different locations. CPA firms with real-time access to information can provide more timely assistance to clients, which can improve efficiency and decision making for clients. The Internet and e-business also provide CPA firms with new promotional opportunities and more efficient research capabilities.

As mentioned earlier, since the rapid evolution of IT has had such a radical impact on the business environment, public accounting firms have responded by applying Internet and e-business in their daily operations. Each year around April, the American Institute of Certified Public Accountants (AICPA) Top Technologies Task Force provides the accounting profession with a list of what it considers to be the current top-10 technology issues (see http://www.toptentechs.com). From Table 1, we can see that the Internet and e-business issues have grown progressively more complex and increasingly urgent during the past few years. On the 1997 list, the Internet and e-commerce were ranked four and seven, respectively; in 1998, they were ranked one and five, respectively. In addition, more complex issues had developed concerning the Internet, such as intranets, private networks, and extranets. The Internet was listed as the number-two issue in 1999; e-business was listed number one and number two for years 2002 and 2001, respectively. In 2002, the Internet

Table 1 Lists of Top-10 Technology Issues Identified by the American Institute of Certified Public Accountants

RANKING	1997	1998	1999	2000	2001	2002
1	Security	Internet, intranets, private networks and extranets	Y2K issues	E-business	Information security and controls	Business and financial reporting applications
2	Image processing	Y2K issues	The Internet	Information security and controls	E-business	Training and technology competency
3	Communications technologies	Security and controls	Information security and controls	Training and technology competency	Electronically based financial reporting (XBRL)	Information security and controls
4	The Internet and public online services	Training and technology competency	Training and technology competency	Disaster recovery	Privacy	Quality of service
5	Training and technology competency	Electronic commerce	Technology management and budgeting	High availability and resiliency of systems	Training and technology competency	Disaster recovery (includes business continuation and contingency planning)
6	The Year 2000	Communications technologies—general	Disaster recovery	Technology management and budgeting	Disaster recovery	Communication technologies—bandwidth
7	Electronic commerce	Telecommuting/virtual office	Virtual offices	Electronically based financial reporting (XBRL)	Qualified information technology personnel	Remote connectivity tools
8	Workflow technology	Mail technology	Privacy	Net issues	Quality of service	Web-based and Web-enabled applications (Internet)
9	Private networks	Portable technology	Electronic money	Virtual offices	Electronic audit trail	Qualified information technology personnel
10	Electronic data interchange	Remote connectivity	Electronic evidence	Privacy	Application service provider (ASP)	Messaging applications (e-mail, faxing, voicemail, instant messaging)

Source: http://www.toptentechs.com, http://www.aicpa.org/pubs/cpaltr/jan2000/ebusines.htm, http://www.aicpa.org/pubs/cpaltr/jan99/top.htm, http://www.aicpa.org/pubs/cpaltr/jan98/intern.htm, http://www.aicpa.org/pubs/jofa/feb97/technol.htm and http://citp.aicpa.org/

is still listed as one of the top-10 technology issues. It is clear that the AICPA highly encourages accounting professionals to learn more about technology and issues related to the Internet and e-business.

Overview to Chapter

This chapter explains how CPA firms have used the Internet and e-business to expand the types of services they offer to clients and to enhance and streamline many of the existing services they provide, such as offering more timely information to clients. We focus on applications in the audit, assurance, financial analysis, and tax areas and discuss how the Internet can be used to access the various types of information that CPA firms need, along with how firms are using the Internet for promotional purposes. With increased use of the Internet, some issues arise, particularly with regard to data security and legal issues, including confidentiality of information. Finally, we explore some possible future developments for firms with respect to the Internet and e-business applications and how they may change the nature of CPA firms' practices.

KEY APPLICATIONS

Enhancing and Expanding Services

New Assurance Services

In late 1990s, the AICPA introduced two new assurance services (WebTrust and SysTrust) to address risks associated with information systems and to enhance systems reliability and e-business security (Boritz, Mackler, & McPhie, 1999; Koreto, 1998a, 1998b). WebTrust is an attest-level engagement provided by specially licensed public accountants to build trust between consumers and companies doing business over the Internet. A WebTrust seal on a company's Web site indicates that the company has a WebTrust audit performed on an average of every 180 days, and that the site complies with the WebTrust principles and criteria in all or part of the four areas: business practices and information privacy, transaction and service integrity, information protection/security and privacy, and availability and controls over certification authorities. CPAs interested in this assurance service niche need to have experience in attestation engagements and knowledge in information technology (for detailed information, see http://www.webtrust.org).

Technically, WebTrust, as the name implies, focuses on Internet-based systems only. On the other hand, SysTrust is an assurance service in which public accountants independently test and verify the reliability of a company's overall system, measured against the essential SysTrust principles: availability, security, integrity, and maintainability. SysTrust plays an important role in conducting e-business, because it is designed to increase the confidence of management, business partners, customers, stockholders, and government agencies in IT and systems that support business operations or any particular activity. Without sufficient confidence in a company's systems, business partners, employees, and external information users may not conduct e-business with the company. Public accountants interested in providing this assurance service must have information systems audit experience. It could be desirable to work with certified information systems auditors (CISAs) on SysTrust engagements. For detailed information, visit http://www.aicpa.org/assurance/systrust/and http://www.systrustservices.com.

Business and Financial Reporting Applications

Applications of technology for business and financial reporting purposes rose to the top of the 2002 top-10 technology issues list after being in third place in 2001. The AICPA has long foreseen the need for improved financial reporting capabilities, and its concern led to the creation of XBRL (extended business reporting language) in 2000. XBRL is an XML-based specification for preparing, distributing, and analyzing financial information (Strand, McGuire, & Watson, 2001). XBRL has been named as the next-generation digital language of business that can ensure the integrity of electronic financial reports (Rezaee, Hoffman, & Marks, 2001). Mike Willis, partner at PricewaterhouseCoopers (PWC) and chairman of the XBRL Project Committee, said, "It is a natural extension of today's Internet technology to the financial reporting process. XBRL provides a platform for the future of business reporting over the Internet, and it will be fundamental to the way companies communicate with stakeholders" (http://www.xbrl.org/). XBRL can standardize financial reporting over the Internet, and companies will be able to use the Web to report information in a timely and accurate manner.

Although XBRL's primary target is commercial and industrial companies for their external financial reporting needs, XBRL can also be used for data analysis, reporting needs, and governmental filings (Hannon, 2001). In 2002, the Securities and Exchange Commission (SEC) created an online repository of XBRL data and financial reports through its EDGAR (electronic data gathering analysis and retrieval) program (Edgar Online, 2002). In addition, XBRL can be implemented for various reporting needs in different industries and different countries. XBRL usage has grown around the world in countries such as Japan, Singapore, Germany, and South Africa (Hannon, 2002). XBRL is only beginning to gain visibility. Because most companies do not know enough about XBRL to understand how it can help their business and how to implement it, there are tremendous opportunities for CPAs to provide consulting services in this area such as choosing proper tools to link a client's internal financial reporting system to the client's Web site for external electronic reporting.

Online Services

The Internet allows CPA firms to interact with clients anytime and anywhere and thus allows for new business models for firms. Possible models include offering clients online audit/review and online consulting services. A few CPA firms have partnered with companies such as Intacct, Oracle Small Business Suite, and Creative Solutions to offer clients web-based accounting systems. With powerful and secure Internet data centers maintained by business partners, auditors have the confidence that clients' accounting systems are well-maintained and have fewer concerns about systems availability/reliability (personal communication, Robert L. Lewis, Jr., and Wendy Bednarz,

Intacct, May 8, 2002). There are significant security concerns for exchanging consulting information between the client and the CPA firms since most information is extremely sensitive, confidential, and damaging if it falls into the wrong hands. Thus, CPA firms must examine carefully whether proper encryption, authentication, and virtual private network technologies are implemented to secure information transmission of their online services.

The reliability and availability of online real-time systems have made continuous audit and assurance possible. According to Rezaee, Sharbatoghlie, Elam, and McMickle (2002, p. 145), continuous audit and assurance is defined as "a comprehensive electronic audit process that enables auditors to provide some degree of assurance on continuous information simultaneously with, or shortly after, the disclosure of the information." Moreover, both researchers and practitioners indicate that real-time financial reporting has necessitated continuous assurance (Elliott, 2002; Rezaee et al., 2002). As Alles, Kogan, and Vasarhelyi (2002) suggested, both WebTrust and SysTrust are continuous assurance services offered by the AICPA to respond to this need. Given the advances in technology, some companies have developed online audit and review tools for auditors to access the client's database and extract data anytime from anywhere for auditing purposes (e.g., Intacct). Since more and more companies are attempting to provide online real-time financial reports, XBRL also plays a critical role in continuous assurance. Clearly, changes in the audit paradigm will continue in order to meet assurance-user needs in the future. The short useful life of operating systems (OS) and applications may become an obstacle to implementing continuous audits, however. By the time a continuous audit tool is beta tested, installed, and implemented, the underlying OS or applications may be upgraded, patched, or replaced, rendering the audit tool inoperable or causing it to report false signals. CPA firms may need to work with their clients to develop long-term strategic plans regarding how to maintain the stability of the clients' OS and key applications to increase the feasibility of continuous audits.

Other than assurance services, some of the larger CPA firms have used the global 24/7 nature of the Internet to broaden clients' access to expertise within the firm. One of the first and probably best known Internet-based consulting services is "Ernie" created by the international accounting firm Ernst & Young (EY) in 1996. Ernie, later renamed Ernst & Young Online Services (n.d.), has been marketed as an online business consultant that provides low-cost access to EY experts in many areas including audit, tax, human resources, strategy, information technology, personal finance, and specified industries. EY routes questions submitted from subscribers to experts throughout its global practice, as needed, and clients typically are promised an answer within 2 days.

Ernie was designed to serve a market of new and small businesses with annual revenues under $250 million that would benefit from having access to outside experts in a variety of areas for which the business could not afford to have its own in-house expertise. It began by charging a fixed monthly fee and has evolved to be free for clients, with capped fees for online questions and charges for certain tools. Key benefits of the online consulting service is the quick turnaround on questions, because businesses today tend to want "just-in-time" information. In addition to answering online questions, EY Online provides a customized homepage for clients, access to a reference library and news items, access to the client's "EY Team," and some online tools for improved decision making.

Tax Applications and Services

Technology has proved to be a tremendous improvement in tax work due to the rote calculations involved, the link between financial records and tax records, and the nature of how tax returns are designed. For the past few decades, tax preparation software has been used to perform various tax functions, such as the calculation of depreciation on assets and printing of W-2 forms from electronic payroll data. Tax preparation software handles such functions as the flowing of data from financial records to tax returns and from tax form to related tax form, as well as error correction. Use of the Internet to enhance tax software applications allows for quicker and easier updates to the software, greater options for data storage, and links to tax information provided by the Internal Revenue Service (IRS) and state tax agencies, as well as by commercial providers of tax research information.

The Internet also allows a CPA firm to access needed tax information directly from a client's Web-based accounting systems. Web-based tax preparation tools allow CPA firms to manage their tax preparation work by tracking the status of return preparation—who is working on a particular return and how far along it is in the process. Web-based tax applications also allow for customized billing based on the detail needed by clients, and they provide access to tax preparation data and process from any location at any time. In addition, clients can easily e-mail data or files to their CPA firm in a more secure and timely manner than using the mail or a fax machine. Finally, many types of tax returns can be filed electronically today. During the 2002 filing season, the IRS reported that about 46 million taxpayers filed their tax return electronically. Also, about 105,000 tax preparers participated in the e-filing program (IRS News Release, IR-2002-53, April 25, 2002, available at Tax Analysts' Tax Notes Today, 2002 TNT 81-19). The benefits of e-filing include getting more accurate information into the IRS databases and quicker refunds.

Accessing Information

CPAs are dependent on information. They need access to the text of accounting pronouncements on generally accepted accounting principles (GAAP), tax research materials, ethics opinions and rules, and general business news and information. The ability to access much of this information on the Internet has greatly improved the efficiency and mobility of CPAs.

Tax Information

For decades, CPA firms relied on a physical library to access tax statutes, regulations, rulings, other tax research information, and tax forms. In the early 1990s, many of

the commercial providers of tax research materials also offered their materials on CD-ROMs. By the late 1990s, the primary commercial providers of tax research materials, such as the Research Institute of America (RIA) and the Commerce Clearing House (CCH), were providing the materials on the Web. Web-based access to tax research materials has several significant advantages over both paper and CD access. For example, providers can update the materials much more quickly and efficiently on the Web than can be done by sending new pages or CDs to customers. Less office space is needed with a Web-based library. Also, CPA firms have 24/7 access to the Web-based research materials and can access them from anywhere without the need to carry around several CDs. Web-based tax research is efficient because links are inserted into the online documents that enable users, for example, to click on a link in a document to see the full text of the cited case (rather than going to the physical library and pulling a book off the shelf). Finally, the search techniques using Web-based materials are superior to what is possible with a paper-based tax research tool.

Over the past several years, commercial providers of tax research materials have put more and more of their materials online, including treatises and journals. Much tax information is also accessible via the Web site of the IRS. In fact, the IRS Web site is one of the most frequently accessed sites, receiving heavy use by both taxpayers and practitioners. During the tax-filing season in early 2002, the IRS Web site had 1.97 billion hits, which was a 28% increase from 2001. On April 15, 2002, alone, there were 78 million hits to the site (IRS News Release, IR-2002-53, April 25, 2002, available at Tax Analysts' Tax Notes Today, 2002 TNT 81-19).

Other Information

Various commercial publishers provide Web-based access to accounting pronouncements, such as financial accounting standards and SEC documents. The Financial Accounting Standards Board ([FASB], 2000) Web site has information on current activity regarding drafting and reviewing new accounting standards and guidance. Copies of the various pronouncements can be ordered from the FASB Web site. Also, the SEC Web site has many types of items that previously could be obtained only by subscribing to a commercial service or by requesting them from the SEC. With the Web, this information is available immediately and at no cost. The SEC Web site provides links to statutes, press releases, special reports and studies, and regulatory actions. In addition, the well-known "EDGAR" service on the SEC Web site (http://www.sec.gov/) allows visitors to view or download copies of the reports (such as 10-Ks) filed by publicly-traded companies.

CPA firms also find value in using the Internet as an information source because many business journals, including the *Wall Street Journal,* can be viewed online, including archives of older articles. CPA firms can also access a variety of information useful in their work at portal Web sites designed specifically for CPAs. For example, the CPAnet.com site provides links to a wide range of accounting and tax news items, articles, conferences, job postings, and even accounting jokes. This portal also includes discussion groups where people can post questions and hope that another member of the discussion group offers an answer. These discussion groups significantly broaden the professional reach of a CPA, although they are not often used to their full potential.

CPA Organization Information

CPA firms can also find a great deal of useful information at Web sites run by accounting organizations, such as AICPA or state societies of CPAs. In March 2000, the AICPA and state CPA societies partnered to launch "CPA2Biz," a service to provide information and products to members (the state societies later left the arrangement). All AICPA products and services (such as registration at AICPA conferences) are only available at the CPA2Biz site. The site offers low-cost access to the AICPA's Resource Online, which enables users to search for documents or view specific accounting reference materials. CPA2Biz also offers online courses (there is a charge for most courses), business application software (such as for payroll and billing), job search and resume posting services, and accounting news and product updates via e-mail (*CPA Insider*). CPA2Biz has relationships with companies that have invested money in CPA2Biz, such as Microsoft (some of the CPA2Biz Web features are only supported by Microsoft's Internet Explorer) and Thomson (members can buy products from this company at the site).

Access and Controversy

CPA2Biz is marketed as a "revolutionary site" and a "single source" that will address all of a CPA's professional needs. The launch of the site caused some controversy between CPA2Biz and some CPA firms, however, primarily because of the AICPA's work in establishing and being an investor in a for-profit venture as well as the fact that AICPA's management received a small ownership percentage (1.6%). The AICPA, its members, and state CPA societies owned 40% of CPA2Biz. In March 2002, AICPA president and chief executive officer, Barry Melancon, announced that he was donating his 1% stock interest in CPA2Biz to the AICPA Foundation. In October 2001, national accounting firm BDO Siedman filed a lawsuit against CPA2Biz. BDO's complaint calls for an injunction based on such causes of action as unfair competition, restraint of trade, and breach of fiduciary duty. The litigation and the degree to which CPAs use the CPA2Biz site will certainly affect the future direction of this for-profit venture designed to provide quick access to products, services, and information for members.

Public Relations

The Internet has provided businesses, including CPA firms, another vehicle to promote themselves. Much of what is on a CPA firm's Web site is similar to what could be in a printed brochure. Many firms have taken advantage of the relatively low cost yet wide reach, of the Internet and provided more information about their firm on the Web than they would place in a brochure. For example, some CPA firms have financial "calculators" available on their Web site (although typically the calculators are not proprietary to the firms) to allow visitors to calculate such things as mortgage payments on a potential home purchase and

how much to save to reach a particular target. The set up of a Web site also enables clients and potential clients to get right to the information they want, even though the company may have a great deal of other, unrelated information on its Web site.

Key promotional items likely to be found at the Web site of many public accounting firms include the following:

- Contact information, firm history, firm mission statement and core values, and biographies of owners and key employees
- Press releases about personnel changes and new activities
- Promotional information—what is so special about the firm, why someone should hire the firm
- Description of services provided, often within industry areas of expertise (such as banking or real estate)
- Downloadable and Web-viewable newsletters and informational reports (such as to explain a tax or accounting rule)
- Tip of the week (typically a tax tip) that may lead clients and potential clients to visit the Web site more often
- Upcoming events, such as seminars
- Career information (types of career opportunities, positions available, how to apply, and ability to submit a resume via e-mail)

Some firms, particularly large international firms, maximize the technology and broad reach of the Internet in ways that go beyond just using the Web as an electronic marketing brochure. Some firms offer free webcasts of technical subjects to their clients and others. Such sessions may involve both a conference call and a Web-based presentation, as well as an option to allow participants to ask questions (either online or via phone). The sessions are typically run by the firm's experts on the particular topic. Because the presenters do not all have to be in the same room (or city), these types of educational programs enable the firms to avoid travel costs, as well as the costs of a room rental for traditional face-to-face seminars. Generally, the participants are offered continuing professional education credits (CPE) for their participation, which is an added incentive to participate. A firm benefits by exposing a large group of clients and nonclients to the firm's experts. Examples of topics covered in Web-based seminars include dealing with new IRS audit and appeals changes (Deloitte & Touche, May 2002) and proper application of Financial Accounting Standard #133 on derivatives (Ernst & Young LLP, December 2001). The presenting firms may also archive the presentations on their Web site for access by anyone at anytime.

As CPA firms expand their use of Web-based accounting tools, it is likely they will offer even more services to clients, such as access to their CPAs' calendars so clients can schedule appointments. Also, firms might set up their Internet services to allow clients to access their own tax returns and other documents prepared for that particular client, assuming that the obvious security concerns can be adequately addressed.

IMPLEMENTATION CONSIDERATIONS

System Reliability and Security

To provide services online, CPA firms must have systems with high reliability and security. System reliability is about a system's availability (available for operations and to be updated and maintained in a manner that continues to provide system availability) and its integrity. A reliable system can operate without material error, fault or failure during a specified time in a specified environment (Boritz et al., 1999).

System security is the ability to protect information resources from unauthorized access, modification, and destruction. Information resources in an Internet/e-business environment are hardware, software, and telecommunications. For CPA firms, online security is vital not only because it is required to protect the information assets, but also because of the long-term trusted relationship with clients. From a client viewpoint, security is the perceived guarantee that no unauthorized parties will have access to communications between the client and the CPA firm. The focus of online security is threefold: authentication, confidentiality, and integrity (Romney & Steinbart, 2000). Authentication is the ability of the system to verify that users are who they claim they are. Confidentiality refers to limiting data access or use to authorized individuals only. Online systems must be able to authenticate the identities of those who attempt to log on, allowing only legitimate users to access the information or database. Integrity refers to maintaining data accuracy and preventing hardware failure and unauthorized tampering. Current encryption technology (128-bit Data Encryption Standard) with public–private key usage and a good public key infrastructure (PKI) can accomplish these three goals. To have a good PKI, a firm needs to form consistent agreement between the practices of a certificate authority (CA) and the firm's certificate policies because the CA manages the firm's public keys.

In addition, properly trained IT professionals can play a key role to make an information system reliable and secure. It is also important that the top management of a firm maintain a well-established system development life cycle policy to assure the reliability and security of its information systems.

Training and Certification

During the 1990s and continuing today, CPA firms are devoting more time to training and education in IT areas, as evidenced by the number of IT conferences, the emergence of IT committees and IT newsletters within state CPA societies, and a new IT certification provided by the AICPA.

One of the premier IT conferences is the annual AICPA Tech conference. In 2002, this 4-day conference consisted of more than 50 sessions within the areas of consulting, technology, products, and IT management. Session topics dealing with the Internet and e-business included securing e-mail, Web collaboration tools, wireless technology and products, Web-based accounting software, e-commerce software systems, SysTrust and WebTrust, maximizing traffic to a Web site, and technology

consulting. Continuing education IT programs offered by state CPA societies include such topics as Web-based financial reporting and analysis (XBRL), security, and expanding a CPA firm's services through use of new technologies.

Many state CPA societies, as well as the AICPA, have IT committees to serve members who specialize in that area and members who want to increase their IT knowledge so they can expand and enhance the services they offer and can assist clients with their IT needs. For example, the Florida Institute of CPAs (n.d.; personal communication, Hue T. Reynolds, April 16, 2002; personal communication, Stam W. Stathis, April 24, 2002) has an E-Commerce Section that provides an Internet-based chat room for members, online expert Q &A, and a member directory. The institute is interested in helping its members expand their use of the Internet beyond just e-mail. The institute and its E-Commerce Section see IT as enabling members to expand their services into e-business opportunities and to share their IT expertise with clients who are seeking to expand their services through use of IT.

The AICPA's Information Technology Section is open to all AICPA members and qualifying non-CPAs. Members receive IT updates and a software news report (eight times per year), as well as Technology Alerts on major technology news. The ability to network with a large group of other CPAs involved with IT work is also a benefit of joining. The section also sponsors the annual AICPA Tech conference.

In 2000, the Information Technology Alliance (ITA) merged into the AICPA to form the IT Alliance. ITA was a 30-year old organization made up of value-added resellers, accounting software vendors, chief information officers (CIOs), chief technology officers (CTOs), and CPAs involved with technology. The ITA members joined AICPA members interested in IT consulting to form the new AICPA section. The IT Alliance existed within the AICPA along with the IT Section until they separated in April 2002 due to strategic decisions of both the ITA and AICPA. The AICPA continues to be an institutional member of the ITA, and the organizations will continue to work together in some ways. The primary focus of each organization varies somewhat in that the IT Section of the AICPA focuses more on assisting CPAs in using technology more effectively, whereas the ITA focuses on assisting members (which includes CPAs) in their roles as providers of IT-based solutions for clients.

Certified Information Technology Professional

In 2000, the AICPA began a new IT certification for CPAs. The designation is known as the Certified Information Technology Professional (CITP). The CITP designation helps the public to view CPAs as IT professionals, improves the quality of the IT services provided by CPAs, and aids in the development of practices in the IT area. A CITP is described by the AICPA as someone who serves in an organization as the "bridge between management and the technologist" (AICPA promotional literature). To become a CITP, a person must be a member in good standing of the AICPA, have a CPA license, pay a fee, submit a written statement of intent to comply with the requirements for reaccredidation and payment of the annual renewal fee, and generate at least 100 points through a combination of experience, lifelong learning (such as continuing education seminars), and examination results.

The type of experience that qualifies and that is covered on the CITP examination falls into the following eight categories:

1. Information technology strategic planning (18%)
2. Information systems management (15%)
3. Systems architecture (11%)
4. Business applications and e-business (16%)
5. Security, privacy, and contingency planning (11%)
6. System development, acquisition, and project management (13%)
7. Systems auditing/internal control (8%)
8. Databases and database management (8%)

The percentages shown indicate the weight given to that topic on the CITP examination. This 2-hour, computer-based exam is administered twice per year and consists of 100 objective questions. To pass, a member must answer at least 75 questions correctly. The CITP Web site provides considerable information on the eight topics, including links to articles on specific technologies, uses, and implementation. The CITP Web site is coordinated with the Top Tech site sponsored by the AICPA that provides background information on technology issues (such as security and disaster recovery), applications (such as data mining and document management), types of technology (such as wireless and authentication), emerging technologies (such as m-commerce and electronic evidence), and case studies (best practices shared by practitioners). The AICPA also offers training to help members earn the CITP designation.

Information Systems Audit and Control Association

Information systems (IS) audits have played an important role in the public accounting profession. Weber (1999) defined IS auditing as the process of collecting and evaluating evidence to determine whether an information system safeguards assets, maintains data integrity, achieves organizational goals effectively, and consumes resources efficiently. Sayana (2002) further stated that the purposes of IS auditing are to evaluate the system, to provide assurances that information in the system is being effectively used, and to make suggestions on how to improve the system. Bagranoff and Vendrzyk (2000) described IS audit practice as "stand alone with very close ties [to financial audit]." IS auditors support financial audits by providing risk assessment services to point out weaknesses that may impact the client's financial statements or impact the business as a whole. IS auditors also offer consulting services such as penetration testing and security diagnostics based on the system weaknesses they find in their audits. Most organizations that support IS auditing are involved with the overall improvement of auditing control objectives to limit organizational risk.

The Information Systems Audit and Control Association (ISACA, n.d.) is the most dominant organization in regard to information systems auditing and has an

aggressive vision: "to be the recognized global leader in IT governance, control and assurance" (http://www.isaca.org). ISACA accomplishes this goal by offering services such as research, setting industry standards, and providing information, education, certification, and professional advocacy.

One of the certifications ISACA oversees is the CISA (Certified Information Systems Auditor). It also operates the IT Governance Institute, believing information technology is no longer simply an enabler of an enterprise's strategy but is also an integral part of the strategy. ISACA has been leading the way by developing "globally applicable information systems auditing and control standards" (http://www.isaca.org).

As Gallegos, Manson, and Allen-Senft (1999, p. 6) indicated, "Technology has impacted the auditing profession in terms of how audits are performed (information capture and analysis, control concerns) and the knowledge required to draw conclusions regarding operational or system effectiveness, efficiency and integrity, and reporting integrity." CPA firms must face the challenges by providing more IT training to their staff so that they can broaden the range of services and effectively deliver those services to their clients.

Legal and Regulatory Issues

Much of the work of CPA firms involves financial data that clients want to protect, as appropriate. Thus, when more and more financial data and client communications about that data are performed or are made available electronically, CPA firms need to understand the technology, as well as the law, to be sure that confidential data is protected by privacy features in their system and their firm's office routines. In addition, a limited confidentiality privilege, added to the federal tax system in 1998, requires that CPA firms be aware of how the confidentiality of protected records is maintained so that clients do not lose any CPA–client privilege that may exist with respect to certain records. Another legal and regulatory concern for some CPA firms involves proper advising of clients subject to SEC rules to be sure that financial information posted to a Web site is properly and timely presented. These key concerns—privacy, confidentiality, and Web posting of financial data—are explained next.

Federal Privacy Law

Many CPA firms are subject to the privacy provisions of the 1999 Gramm–Leach–Bliley (GLB) Act. The privacy provisions apply to a broad range of financial services that includes preparation of nonbusiness tax returns and financial and tax planning. The act prohibits those subject to it from disclosing nonpublic personal information without authorization. The act also directs the Federal Trade Commission (FTC) to issue regulations on the disclosure required by companies subject to the privacy provisions. The FTC (2000) issued final regulations in May 2000 and CPAs had to be in compliance by July 1, 2001.

CPA firms subject to the FTC regulations must provide a disclosure notice to new clients and an annual disclosure to all clients that accurately depicts the firm's privacy policy. The disclosure must explain the firm's practices and policies regarding privacy, including such items as the categories of nonpublic personal information collected and other data the firm might disclose, the client's right to opt out of any disclosures by the firm, and how a client's nonpublic personal information is maintained in a secure and confidential manner. The AICPA Web site provides members with information about complying with the act, including sample disclosure letters that can be sent to clients.

The new disclosure rules are most relevant to a CPA in terms of the notice requirement. CPAs are already subject to disclosure and confidentiality rules by their licensing state, the AICPA, and the federal tax law. For example, Rule 301 of the AICPA Code of Professional Conduct states that "a member in public practice shall not disclose any confidential client information without the specific consent of the client." Internal Revenue Code (IRC) section 6713 imposes a penalty on any tax return preparer who discloses information provided to him or her for return preparation or uses such information for any purpose other than to prepare or assist in preparing a tax return. IRC section 7216 provides that such disclosure is a misdemeanor if the disclosure is done recklessly or knowingly.

Confidentiality Privilege

In 1998, the IRS Restructuring and Reform Act created a limited confidentiality privilege for clients of CPAs. This new provision (IRC section 7525) extends the common law attorney-client privilege of confidentiality with respect to tax advice to any federally authorized tax practitioner (attorneys, CPAs, enrolled agents, and enrolled actuaries). This privilege is intended to apply to the same extent as it would between a taxpayer and an attorney; however, it does not expand the attorney–client privilege.

The section 7525 privilege, if otherwise applicable, applies to *tax advice* furnished to a client-taxpayer or potential client-taxpayer. However, the privilege may only be asserted in a noncriminal tax matter before the IRS and any noncriminal tax proceeding in federal court by or against the U.S. "Tax advice" is defined as advice given by an individual with respect to a matter within the scope of the individual's authority to practice as a federally authorized tax practitioner (per Treasury Department Circular 230) that involve matters under the IRC. Thus, the section 7525 privilege cannot be asserted to prevent any other regulatory agency (such as the SEC) or person from compelling the disclosure of information. The section 7525 privilege does not apply to any written communication between a federally authorized tax practitioner and a director, shareholder, officer, or employee, agent, or representative of a corporation in connection with the promotion of the direct or indirect participation of the corporation in any tax shelter (per the definition at IRC section 6662(d)(2)(C)(iii)). CPAs need to check their state's law to see if the state has conformed to the federal privilege.

Section 7525 goes beyond Rule 301, Confidential Client Information, of the AICPA Code of Professional Conduct (noted earlier), because the section 7525 privilege is legally enforceable and generally will prevent disclosure, even if compelled by the IRS through a summons.

The existence of a CPA-client privilege means that CPAs need to understand the basics of the attorney–client

privilege as well as the limitations of section 7525. CPAs need to know what is considered a confidential communication and what types of tax work and documents are protectable. In addition, CPA firms will need to implement office practices to be sure that no disclosure occurs that may cause a client's privilege to be waived. For example, inadequate electronic storage or security over the storage may indicate that no confidentiality was intended, thus the information is not privileged. In addition, CPA firms need to evaluate whether any encryption or other precautions are needed to ensure that electronic transmissions protect confidential information.

Questions have been raised over the past several years by attorneys and bar associations as to whether e-mail is a confidential delivery vehicle such that information sent via e-mail is privileged (assuming it otherwise qualifies for protection under the privilege). Questions have also arisen as to whether certain rules, such as those dealing with solicitation, apply to e-mails and information provided on Web sites. Some states have issued guidance on these matters. For example, in 1997, the Illinois State Bar Association issued Advisory Opinion No. 96–10. The conclusion reached is that an attorney's duty to protect confidential client information is not violated by the attorney's use of e-mail and the Internet without encryption to communicate with clients unless unusual circumstances require enhanced security measures (such as when it is already known that break-ins have been attempted). The rationale is that the ability to intercept e-mail is about as difficult as intercepting a regular phone call. Also, intercepting e-mail is illegal under the Electronics Communications Privacy Act of 1986. Before communicating via e-mail with a client or potential client, however, consideration should be given to who else has access to the e-mail. For example, if the client is using the e-mail system at his or her job site and it is regularly reviewed by the systems administration staff or is shared e-mail, there is no expectation of privacy and thus no indication that the communication was intended to be confidential. (Also see American Bar Association Formal Opinion No. 99–413.)

There is little case law on the subject of e-mail and confidentiality, and no guidance with respect to the IRC section 7525 privilege. Several states, however, have issued opinions similar to that in Illinois, which may provide some general guidance for a CPA. CPA firms will need to consider the limited guidance that exists, the basics of the privilege, and the nature of the information involved and the security situation (for example, is it one prone to hackers?) in establishing the procedures needed to maintain the client's privilege under section 7525 with respect to electronic transmission of protected tax information. Future guidance from the IRS may provide some assistance on this matter as well.

Privacy in Practice

A CPA firm doing business over the Internet, such as online consulting, will need to demonstrate to clients that information transferred electronically and Web-based accounting information is secure from people who are not supposed to access it and view it. Clients will also need to know that the CPA firm's data storage systems are secure. Basically, to be successful and to operate within a CPA's professional responsibility, the CPA firm may want to provide the same key protections provided by a seal of approval such as CPA WebTrust (see earlier discussion of this service). The three assurances offered by WebTrust are proper disclosure of business practices for e-business transactions, integrity of transactions, and protection of information. In essence, CPA firms will certainly find that clients will want the same privacy and security protections from the online services they receive from their CPA that their CPA, as a trusted business advisor, should be recommending for the client's business.

Regulatory Considerations in Online Financial Reporting. Since 1995, the SEC has issued various releases providing guidance to companies and their financial advisors on procedures to allow for electronic delivery of financial information. Generally, use of technology to deliver information is encouraged because of its efficiency in allowing for quick and wide distribution of information in a cost-efficient manner. The 1995 and 1996 SEC releases provide guidance to ensure that the electronically delivered information is at least equivalent to paper delivery (Securities Act Releases Nos. 7233 [1995] and 7288 [1996]). The creation of XBRL has made use of the Internet to deliver and present all types of financial information in a standardized language, a reality. The usage of electronic financial reporting will continue to result in the need for more guidance to ensure that the information is as complete and reliable, however, as has been expected with paper disclosures. Securities Act Release No. 7856 (2000) addressed issues that can arise when a registered company's Web site includes links to Web sites of third parties that include financial information about the company. For example, to what extent is the company liable under the antifraud provisions of the securities laws for the financial information at the third party's Web site? The release states that the answer depends on the facts and circumstances of the particular situation. Three factors to be considered are the context of the hyperlink, the risk of investor confusion, and the presentation of the hyperlinked information. (View Internet-related SEC Interpretive Releases at http://www.sec.gov/divisions/enforce/internetenforce/interpreleases.shtml.)

Taxation. In 1998, the federal government enacted the Internet Tax Freedom Act (1998) providing a three-year moratorium prohibiting state and local governments from imposing certain taxes on Internet access fees. The Internet Tax Nondiscrimination Act (2001) extended the moratorium to November 1, 2003. Debate continues on how certain taxes should apply to Internet transactions. Existing tax rules were not written with the e-commerce business model in mind, and sometimes online transactions do not fit neatly within existing tax rules, and uncertainty remains. Tax issues also exist in that some policymakers believe that certain Internet transactions should not be taxed so that the Internet and e-commerce will flourish.

CPA firms get involved in the e-commerce taxation debates and issues because of their expertise with tax rules and their obligation to advise clients. The area that has received the most attention involves the application of sales

tax to e-commerce transactions. CPA firms advising businesses setting up e-commerce operations need to be aware of the existing rules governing e-commerce taxation to ensure that their clients have proper procedures in place to collect any tax owed and can structure their e-commerce operations to obtain the best tax planning results. Given the complexity of some of the issues and their often global nature, CPAs will need to stay current on developments in this area so they can properly advise their clients.

FUTURE POTENTIAL—TRENDS AND OPPORTUNITIES

The rapid development of IT has significantly changed the business environment and business models and processes; hence, the accounting profession must respond to the new challenges and take the opportunities to broaden its service spectrum. Because assurance services are critical to all business reporting and the Internet has made continuous reporting possible, CPA firms now have many opportunities to perform continuous audit and assurance. Continuous reporting is real-time reporting, meaning that digitized information becomes available through electronic channels simultaneously with its creation (Elliott, 2002). Many issues are involved with such a practice. Because various types of information often flow to creditors, investors, trading partners, government agencies, and employees, it is vital that the systems provide reliable information (e.g., SysTrust assurance service), that a company maintains its Web sites in a way that the external users of the information can trust it (e.g., WebTrust assurance service), and that the information provided is easy to be downloaded for analysis purposes (i.e, XBRL). Therefore, continuous audit and assurance will be the trend for the accounting profession (both external and internal audit practices).

In addition, the Internet leads to a more global reach of CPA firms, which may cause concern over national or international licensing standards. For this reason, CPA firms may want to encourage more staff to obtain the CISA (Certified Information Systems Auditor) certification because it is recognized internationally. Most important, the business world is moving toward a paperless, or even a virtual, office in which all records are stored on the Web and accessible anywhere and anytime. Because the information is so easily accessed, CPA firms must be especially careful to make their staff aware of all the related legal and regulatory issues, such as privacy and confidentiality of clients' online information, when the firm is using that information for audit or consulting engagements.

Finally, we will likely see CPA firms continue to find innovative ways to use the Internet and e-business opportunities to further expand and enhance their services and enable companies to provide more information and services to their clients.

GLOSSARY

Certified public accountant (CPA) Accountants licensed by a state agency to perform certified financial audits of businesses and other organizations. CPAs typically must have a certain number of university- or college-level courses in accounting and related business subjects and a number of hours of experience. In addition, they must pass a national examination.

Encryption The conversion of data into a secret code for transmission over a public network. The original text, or "plaintext," is converted into a coded equivalent called "ciphertext" via an encryption algorithm. The ciphertext is decoded (decrypted) at the receiving end and turned back into plaintext.

Extensible markup language (XML) An open standard for describing and defining data elements on a Web page and business-to-business documents. It uses a similar tag structure as HTML; however, whereas HTML defines how elements are displayed, XML defines what those elements contain. HTML uses predefined tags, but XML allows tags to be defined by the developer of the page. Thus, virtually any data items, such as products, sales representative's name, and amount due, can be identified, allowing Web pages to function like database records. By providing a common method for identifying data, XML supports business-to-business transactions and is expected to become the dominant format for electronic data interchange.

Electronic data gathering analysis and retrieval (EDGAR) A reporting system that public companies must use to send financial data to the Securities and Exchange Commission. In late 1990s, EDGAR was revamped to accept HTML and PDF files.

CROSS REFERENCES

See *Extensible Markup Language (XML); Taxation Issues; XBRL (Extensible Business Reporting Language): Business Reporting with XML.*

REFERENCES

AICPA CPA2Biz information. Retrieved March 3, 2003, from http://www.cpa2biz.com/

AICPA Information Technology Section. Retrieved May 17, 2002, from http://www.aicpa.org/members/div/infotech/index.htm

AICPA Top Tech Issues. Retrieved May 17, 2002, from http://www.toptentechs.com/

Alles, M., Kogan, G. A., & Vasarhelyi, M. A. (2002). Feasibility and economics of continuous assurance. *Auditing: A Journal of Practice & Theory, 21,* 125–138.

Awad, E. M. (2002). *Electronic commerce: From vision to fulfillment*. Upper Saddle River, NJ: Prentice Hall.

Bagranoff, N. A., & Vendrzyk, V. P. (2000). The changing role of IS audit among the Big Five US-based accounting firms. *Information Systems Control Journal, 5,* 33–37.

Boritz, E., Mackler, E., & McPhie, D. (1999). Reporting on systems reliability. *Journal of Accountancy, 186,* 75–87.

Deitel, H. M., Deitel, P. J., & Steinhuhler, K. (2001). *E-business and e-commerce for managers*. Upper Saddle River, NJ: Prentice Hall.

EDGAR Online. (2002). XBRL: How it can improve today's business environment. In *XBRL Express*. Retrieved May 23, 2002, from http://www.EDGAR-online.com/XBRL/XBLR_today.asp

Electronics Communications Privacy Act of 1986, Title 18, U.S.C. §2510 et Seq. (1986).
Elliott, R. K. (2002). Twenty-first century assurance. *Auditing: A Journal of Practice & Theory, 21,* 139–146.
Ernst & Young Online Services. (n.d.) Retrieved May 25, 2002, from http://eyonline.ey.com
Federal Trade Commission. (2000, May 24). Final regulations on privacy of consumer financial information. *Federal Register, 65*(101), 33688.
Financial Accounting Standards Board. (2000). *Electronic distribution of business reporting information*. Retrieved May 25, 2002, from http://www.fasb.org/brrp/brrp1.shtml
Florida Institute of CPAs. (n.d.). Retrieved May 1, 2002, from http://www.ficpa.org/
Gallegos, F., Manson, D. P., & Allen-Senft, S. (1999). *Information Technology Control and Audit*. Boca Raton, FL: Auerbach CRC Press.
Gramm–Leach–Bliley Act. Pub. L. No. 106–102 (1999).
Greenstein, M., & Vasarhelyi, M. (2002). *Electronic commerce: Security, risk management, and control*. New York: McGraw-Hill Irwin.
Hannon, N. (2001). XBRL: Not just for financial statements anymore. *Strategic Finance, 83,* 65–66.
Hannon, N. (2002). XBRL makes progress worldwide. *Strategic Finance, 83,* 61–62.
Information Systems Audit and Control Association. (n.d.). Information Systems Audit and Control Association and Foundation. Retrieved May 23, 2002, from http://www.isaca.org
The Internet Tax Freedom Act, Title XI of Pub. L. No. 105-277 (1998).
Internet Tax Nondiscrimination Act, Pub. L. No. 107-75 (2001).
IRS Restructuring and Reform Act of 1998. Pub. L. No. 105–206 (1998).
Kogan, A., Sudit, E. F., & Vasarhelyi, M. A. (1998). *The internet guide for accountants*. Upper Saddle River, NJ: Prentice Hall.
Koreto, R. J. (1998a). WebTrust: A new approach to e-commerce. *Journal of Accountancy, 185,* 38.
Koreto, R. J. (1998b). A WebTrust experience. *Journal of Accountancy, 185,* 99–102.
Rezaee, Z., Hoffman, C., & Marks, C. (2001). XBRL: Standardized electronic financial reporting. *The Internal Auditor, 58,* 46–51.
Rezaee, Z., Sharbatoghlie, A., Elam, R., & McMickle, P. L. (2002). Continuous auditing: Building automated auditing capability. *Auditing: A Journal of Practice & Theory, 21,* 147–163.
Romney, M. B., & Steinbart, P. J. (2000).*Accounting information systems*. Upper Saddle River, NJ: Prentice Hall.
Sayana, S. A. (2002). The IS audit process. *Information Systems Control Journal, 1,* 20–21.
SEC Securities Act. Releases, testimony and other reports on reporting financial data online. Retrieved May 17, 2002, from http://www.sec.gov
Strand, C., McGuire, B., & Watson, L. (2001). The XBRL potential. *Strategic Finance, 82,* 58–63.
Weber, R. (1999). *Information systems control and audit*. Upper Saddle River, NJ: Prentice Hall.
XBRL. (2002). Retrieved May 22, 2002, from http://www.xbrl.org/

FURTHER READING

AICPA. (1998). Top 10 technologies stress communications. *Journal of Accountancy, 185,* 22–23.
AICPA. (1999). Y2K tops tech issues list. *Journal of Accountancy, 186,* 16–17.
Harding, W. E., & Zarowin, S. (2000). Finally, business talks the same language. *Journal of Accountancy, 187,* 24–30.
Information Technology Alliance. Retrieved May 26, 2002, from http://www.italliance.com
Intacct. (n.d.). Retrieved May 24, 2002, from http://www.intacct.com/
Ratliff, R. L., Wallace, W. A., Sumners, G. E., McFarland, W. G., & Loebbecke, J. K. (1996). *Internal auditing principles and techniques* (2nd ed). FL: Institute of Internal Auditors .
Smith, S. (1997). *Top 10 technologies and their impact on CPAs*. AICPA Technology Series, New York: AICPA.
Tie, R. (2000). E-business Top Tech Priorities for CPAs. *Journal of Accountancy, 189,* 20–21.

Public Key Infrastructure (PKI)

Russ Housley, *Vigil Security, LLC*

INTRODUCTION

As more business transaction occur on the Internet, security services based on cryptography become essential. Public key cryptography plays an important role in providing confidentiality, integrity, authentication, and nonrepudiation. The basic problem with using public key cryptography is determining who holds the corresponding private key. There are two ways to address this problem. In the first approach, the public key user maintains a local database of the public key and identity pairs. This approach is used in secure shell (SSH) and account-based secure payment as defined in ANSI X9.59, but it does not scale to large communities or facilitate ad hoc communications. The second approach does not have these shortcomings. In the second approach, a trusted party issues a *public key certificate,* or simply *certificate,* containing identification information and a public key. The recipient of such a certificate can be confident that the named party has possession of the private key that goes with the public key contained in the certificate. The collection of hardware, software, people, policies, and procedures needed to create, manage, store, distribute, and revoke certificates is called a *public key infrastructure* (PKI).

The certificate may also indicate the applications that it supports. A certificate *issuer,* called a *certification authority* (CA) can specify the supported applications or specify the expected cryptographic operations. For example, the certificate could specify virtual private network (VPN) key management. Alternatively, the certificate issuer might specify that the public key should be used for validating digital signatures.

PKI is not an application in its own right; rather, it is a pervasive substrate. When properly implemented, it can be taken for granted. PKI provides the binding of public keys and identity information, and then applications make use of the public keys to provide security services such as confidentiality, integrity, authentication, and nonrepudiation.

PKI BASICS

The public key certificate contains fields for the subject's identity and public key. The certificate can indicate a company or organization along with a common name. A variety of name forms are supported. Some name forms are abstract, and others are addresses, such as an e-mail address. The certificate also includes two date fields that specify an activation date and an expiration date. The certificate also contains the name of the CA that created the certificate. To clearly identify each certificate that it issues, the CA includes a unique serial number. Finally, the entire contents of the certificate are protected by the CA's digital signature. Figure 1 illustrates Bob's public key certificate.

In Figure 1, the Hawk CA1 issued Bob's public key certificate. The certificate was activated at noon on February 14, 2002, and will expire at noon on February 14, 2003. This certificate has serial number 48. It includes Bob's name and his RSA public key. The Hawk CA1 signed the certificate with it's own private key, using the DSA signature algorithm and the SHA-1 one-way hash function.

The CA's signature ensures that the certificate cannot be undetectably modified. If anyone changes the contents of the signed certificate, it can be easily detected. The signature will not validate with the modified certificate content. If the digital signature does not verify, the contents have been changed or the certificate is a complete forgery. Either way, it will not be trusted.

How can a certificate user determine whether to trust the certificate contents? The certificate cannot indicate whether the subject has died or changed jobs. Similarly, by looking at a credit card, merchant cannot tell whether it has been revoked.

Like business cards, once a certificate is distributed, it is practically impossible to retrieve all of the copies. In fact, the problem is worse for certificates, since they are digital objects, certificates can be easily replicated and redistributed. All copies cannot be recovered if the information in it is no longer current.

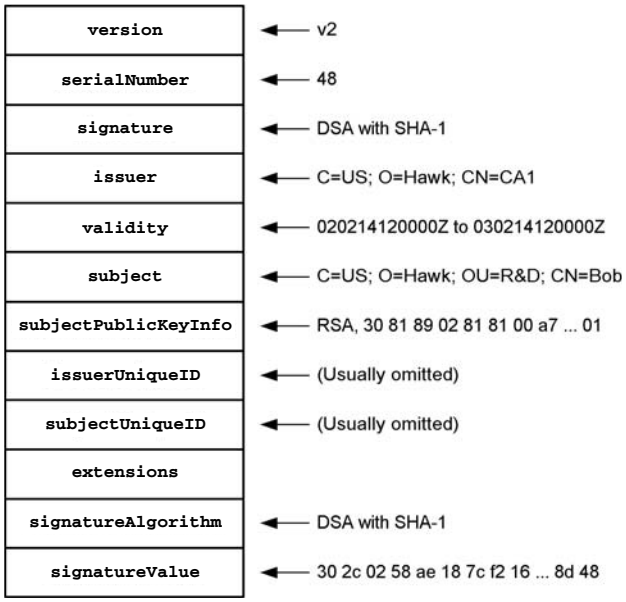


Figure 1: X.509 certificate structure.

The CA's job is to link a public key with a subject's identity in a trustworthy fashion. If the subject notifies the CA that the certificate is no longer correct, then the issuer needs to get that information to anyone who uses the certificate. To determine if the certificate is still trustworthy, the certificate user supplements the unexpired certificate with additional information: either a *certificate revocation list (CRL)* or an *online certificate status protocol (OCSP)* response.

The CRL contains a digitally signed list of serial numbers from unexpired certificates that should not be trusted. The CA generates a CRL regularly and posts it for anyone to obtain. The CA includes the issuance date in the CRL, and usually a date by which an updated CRL will be published. This allows the certificate user to be sure that current information is used. Figure 2 illustrates a CRL that revokes Bob's certificate.

Alice would like to determine the status of Bob's certificate, so she obtains a CRL issued by Hawk CA1. The CRL was issued at 6:00 p.m. on April 15, 2002, and the next issue can be expected 24 hours later. The CRL includes a list of certificate serial numbers for *revoked* certificates.

Alternatively, the OCSP Response provides the revocation status for a single certificate. The certificate user sends a query to a trusted server using the OCSP, suspending acceptance of the certificate in question until the server returns a digitally signed response. In some circumstances, OCSP can provide more timely revocation information than CRLs. More important to many applications, OCSP can also provide additional certificate status information.

One CA cannot reasonably issue certificates to every Internet user. Obviously, there will be more than just one. It is not possible for every Internet user to investigate each CA and determine whether the issuer ought to be trusted. A company might provide certificates to its employees; a business might provide certificates to its customers and business partners; or an Internet user might select a CA

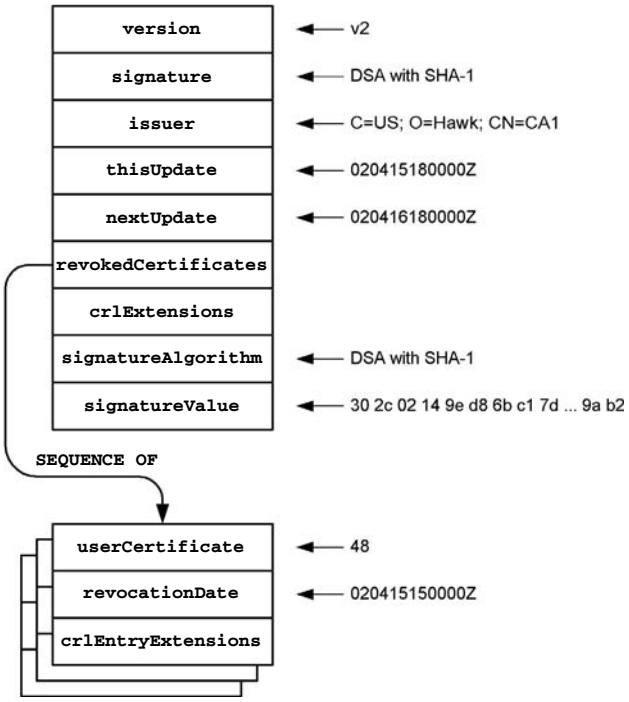


Figure 2: X.509 CRL structure.

to issue their certificate. There are many potential sources of certificate, each satisfying different marketplace needs.

Alice may get a certificate from Hawk CA1. Alice can also trust other CAs that Hawk CA1 trusts. Hawk CA1 indicates this trust by issuing them a certificate. These CAs can indicate trust in other CAs by issuing them certificates. Alice can develop a chain of certificates and automatically decide if certificates from another issuer may be used for the intended purpose. Figure 3 illustrates two

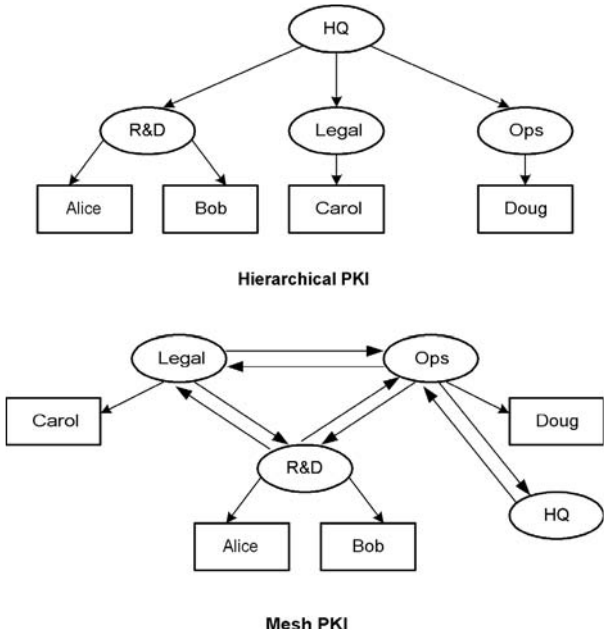


Figure 3: Hierarchical PKI and mesh PKI architectures.

popular PKI construction topologies: the hierarchical PKI and the mesh PKI.

PKI Components and Users

A PKI is often built from three basic functional components: the certification authority (CA), the registration authority (RA), and the repository. A CA is the primary component of the PKI, known by its name and its public key. A CA comprises hardware, software, and the people who operate it. A CA issues certificates, maintains certificate status information and issues CRLs, and publishes certificates and CRLs. A CA must protect the private key or keys used to sign certificates and CRLs, using physical, procedural, and technical controls.

An RA verifies certificate contents prior to certificate issuance, and it may also assume some of the responsibility for certificate revocation decisions. Certificate contents may be verified by information presented to the RA, such as a driver's license. They may also reflect data from a company's human resources department. A CA is likely to work with multiple RAs, because different RAs may be needed for different users groups.

A repository distributes certificates and CRLs. It accepts certificates and CRLs from one or more CAs and makes them available to parties that need them, and it is usually designed to maximize performance and availability. Repositories are often duplicated to maximize availability, increase performance, and add redundancy.

A PKI supports two types of users: certificate holders and relying parties. A certificate holder is the subject of certificate, and it holds the corresponding private key. The CA issues a certificate to the certificate holder. In many circumstances, the certificate holder requests the certificate directly from the CA or through the RA. Certificate holders may need to interact with the repository to obtain their own certificate but do not regularly interact with it. Certificate holders may include their own certificate in transactions.

A relying party uses the public key in a certificate to verify signatures, encrypt data (key transport), or perform key agreement. A relying party identifies one or more trust anchor, verifies signatures on certificates and CRLs, obtains certificates and CRLs from a repository, and constructs and validates certification paths. A relying party regularly interacts with repositories, but it has no interactions with RAs.

PKI ARCHITECTURES

The most basic PKI architecture is a single CA that provides all the certificates and CRLs for a community of users. In this configuration, all users trust the CA that issued their certificate. By definition, new CAs cannot be added to the PKI, and all certificates are user certificates. The users accept only certificates and CRLs issued by their CA. Although the simplest to implement, this architecture does not scale easily to support large or diverse user communities. The single CA PKI presents a single point of failure. Compromise of the CA invalidates the trust point information and all certificates that have been issued in this PKI. Every user in the PKI must be informed about the compromise immediately, or they may establish security based on unreliable information. To reestablish the CA, all certificates must be reissued and the new trust point information must be distributed to all the users. To overcome these deficiencies, two architectures are widely employed: the hierarchical PKI and the mesh PKI. (Recall that Figure 3 illustrates these topologies.)

Hierarchical PKI

The hierarchical PKI is the traditional PKI architecture. All users trust the same central *root CA*. With the exception of the root CA, all of the CAs have a single superior CA. CAs may have subordinate CAs or issue certificates to users or both. A single certificate represents each trust relationship, making certification path construction simple, obvious, and deterministic. The certification paths are usually short. The longest path is equal to the depth of the tree.

Superior CAs may impose restrictions upon the subordinate's actions. These restrictions could be maintained through procedural mechanisms or imposed through the certificates themselves. In the latter case, the CA certificate will contain additional information to describe the restrictions. For example, the Hawk HQ CA could issue a certificate to a subordinate Hawk Legal CA that requires valid certificates to contain a particular prefix in all subject names, which clearly indicates employment in the legal department.

When users are portioned into smaller groups, each served by a different CA in the hierarchical PKI, it is easily handle the compromise of a single CA, as long as it is not the root CA. If a CA is compromised, its superior CA simply revokes its certificate. Once the CA has been reestablished, it issues new certificates to all of its users. The superior issues a new certificate to the CA, containing the new public key, bringing it back into the hierarchy. During the interim, transactions between any two users outside the compromised part of the PKI can proceed. Of course, users in the compromised part of the hierarchy lose all services.

On the other hand, the compromise of the root CA has the same impact as in the single CA architecture. It is critical to inform all the users in the hierarchical PKI that the root CA has been compromised. Until the root CA is reestablished, issues new certificates to its subordinates, and distributes the new trust point information, users cannot use the PKI to establish secure communications. In comparison to the compromise of the single CA, the root CA will have to reissue a much smaller number of certificates to resume operations. The root CA usually operates offline, significantly reducing the likelihood of such a compromise.

Mesh PKI

The mesh PKI architecture is the primary alternative to a hierarchy. Multiple CAs provide PKI services, but the CAs are related through peer-to-peer relationships. Each user trusts a single CA; however, the trusted CA is not the same for all users. Generally, users trust the CA that issued their certificate. CAs issue certificates to each other; a pair of certificates describes their bidirectional trust

relationship. The same constraint mechanisms that were used in the hierarchical PKI may be used to avoid placing unrestrained trust in other CAs.

A new CA can easily be added. The new CA issues a certificate to at least one CA that is already a member of the mesh, who also issues a certificate to the new CA. Path construction is particularly difficult in a mesh PKI; however, it is nondeterministic. Path discovery is more difficult because there are often multiple choices. Some of these choices lead to a valid path, but others result in a useless dead end that does not terminate at a trust anchor. Even worse, it is possible to construct an endless loop of certificates.

Certificates issued to CAs in a mesh PKI are also more complex than the ones usually found in a hierarchical PKI. Because the CAs have peer-to-peer relationships, the certificates contain constraints to control certification paths that will be considered valid. If a CA wishes to limit the trust, it must specify these constraints as certificate extensions in the certificates issued to all of its peers.

Because Mesh PKIs include multiple trust points, they are very resilient. Compromise of a single CA cannot bring down the entire PKI. CAs that issued certificates to the compromised CA simply revoke them, thereby removing the compromised CA from the PKI. Users associated with other CAs will still have a valid trust point and can communicate securely with the remaining users in their PKI. In the best case, the PKI shrinks by a single CA and its associated user community. At worst, the PKI fragments into several smaller PKIs. Recovery from a compromise is simple and straightforward, primarily because it affects fewer users.

Hybrid PKI Architectures

Two approaches are commonly used to join two or more enterprise PKIs: cross-certification and a bridge CA.

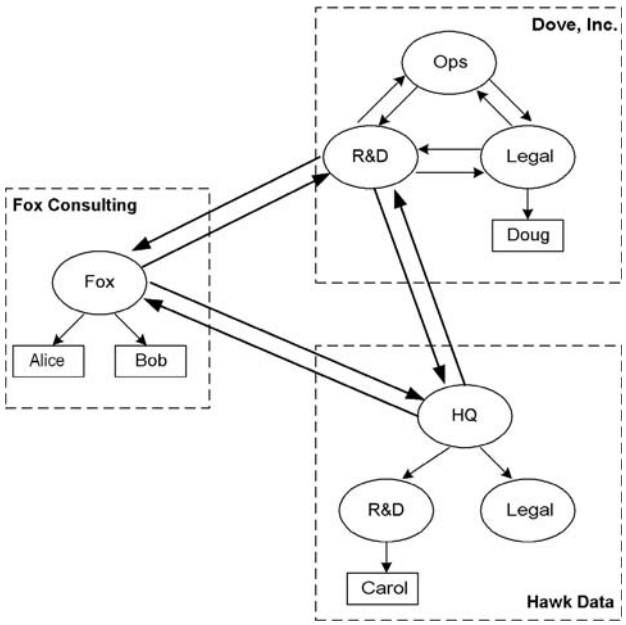


Figure 4: Certification paths with cross certified PKIs.

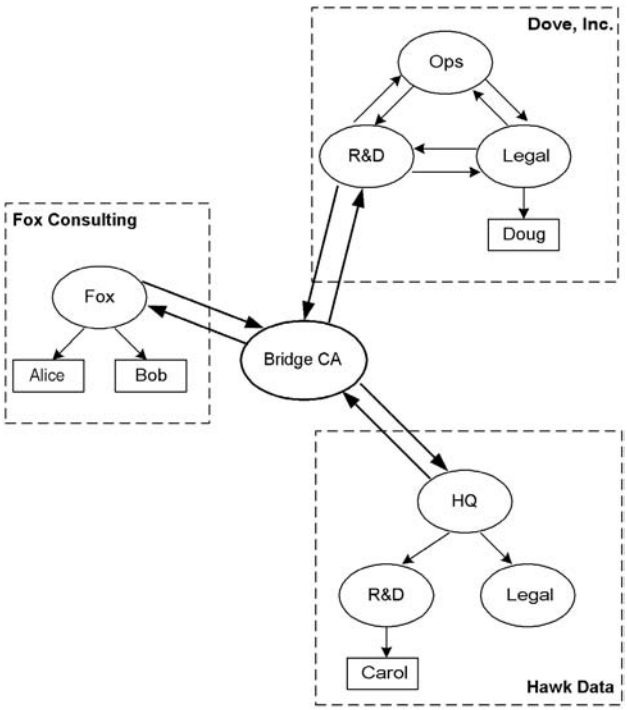


Figure 5: Certification paths with a bridge CA.

Both techniques establish peer-to-peer trust relationships. Figure 4 shows one example of a cross-certified PKI. This architecture is an appropriate solution to establish trust relationships between a few enterprise PKIs. In Figure 4, three peer-to-peer relationships and six CA certificates were required to establish these relationships. This number grows rapidly, however, as the number of enterprise PKIs increases. Cross-certifying n enterprise PKIs requires $(n^2 - n)/2$ peer-to-peer relationships and $(n^2 - n)$ certificates. Establishing these relationships requires a time-consuming review of policies and practices.

Figure 5 shows the same enterprise PKIs establishing trust via a bridge CA. Unlike a mesh CA, the bridge CA does not issue end-entity certificates. Unlike a root CA in a hierarchy, the bridge CA is not intended for use as a trust point. All PKI users consider the bridge CA as an intermediary. The trust relationships between the bridge CA and the principal CAs are all peer-to-peer. It is easy to add new CAs, or entire enterprise PKIs, to a bridge-connected PKI. The change is transparent to the users, because no change in trust points is required. In general, the use of the Bridge CA will require less time to be spent reviewing policies and practices than a comparable Cross-Certified PKI.

Neither cross-certification nor the bridge CA simplified certification path construction or validation. In general, path construction is just as complex as in a mesh PKI; however, path construction can be greatly simplified if CAs are aligned with the name space in which the certificates are issued.

PUBLIC KEY CERTIFICATES

The X.509 public key certificate is named after the document in which it was originally specified: *CCITT Recommendation X.509*. This document, first published

in 1988, specifies the authentication framework for the X.500 Directory. The X.500 Directory requires strong authentication to ensure that only authorized users make modifications. In addition, when the directory contains confidential information, authentication can be used to control directory access.

Over time, the focus shifted from supporting the directory to developing a general-purpose PKI. As a result, two upwardly compatible versions have been published since 1988. Version 2 certificates addressed a single issue: reuse of names. The Version 2 enhancements are rarely used today. Version 3 of the X.509 certificate introduces certificate extensions. Extensions are used when the issuer wishes to include information not supported by the basic certificate fields. All modern PKI implementations generate and process X.509 version 3 (X.509 v3) certificates. The set of extensions used by implementations varies widely.

The Internet Engineering Task Force (IETF) profiled X.509 certificates for the Internet. Like all Internet standards, it is published in a request for comment (RFC) document. The Internet Certificate and CRL Profile, RFC 2459, was published in March 1999. In April 2002, RFC 2459 was replaced by RFC 3280, which identifies optional features of X.509 that are required for the Internet, and it discourages the use of other features.

Subjects and issuers are identified using the distinguished name (DN), a structured type that supports the X.500 hierarchical naming system. The X.500 suite of standards was expected to result in a global directory. This lofty goal required a name form that could be used to create globally unique names. Naming authorities manage their own name spaces, and only that authority assigns names in that space, ensuring collision-free names.

Additional name forms are supported through subject alternative name extension and the issuer alternative name extension. The additional name forms include, but are not limited to, the following:

- Internet domain names (often called DNS names)
- RFC 822 e-mail addresses
- X.400 e-mail addresses
- World Wide Web URLs

CERTIFICATE REVOCATION

Two approaches are used today for certificate status: CRL and OCSP. The basic mechanism for certificate status is the CRL, which is profiled for Internet use in RFC 3280. A CA revokes a certificate by placing the certificate serial number and revocation date on the signed CRL. Certificate users simply search the most recent CRL to determine whether a particular certificate is revoked.

OCSP is specified in RFC 2560, and it enables applications to determine the status of a particular certificate by querying an OCSP responder. A certificate user sends a status request to the OCSP responder, and then the OCSP responder replies with digitally signed certificate status information. The CA can host this service locally, or the CA can delegate this responsibility to an independent OCSP responder.

When using CRLs, the CA name and certificate serial number identify a certificate, but OCSP uses the more complicated certificate identifier. In the absence of a global directory system, it is possible that two CAs could choose the same name. Because an OCSP responder may provide service for multiple CAs, the OCSP responder must be able to distinguish CAs with the same name. Two CAs will not have the same public key, so a hash of the issuer public key is used in addition to the hash of the CA name to identify the issuer.

OCSP is often described as providing revocation information in a more timely fashion than CRLs. An OCSP responder can provide the most up-to-date information it possesses without repository latency. If the OCSP responder is also the CA, the most up-to-date information will be provided. With CRLs, the CA may have additional information that it cannot provide to certificate users. In practice, however, there has been little difference in freshness of the certificate status information provided by an OCSP responder and a CRL.

Most OCSP responders are not CAs. Rather, they are single-purpose machines that handle certificate status requests for a large number of CAs. Typically, these servers obtain their revocation information periodically in the form of CRLs. The information obtained by the requester is no fresher than if they obtained the same CRLs themselves.

The certificate user must place irrevocable trust in the OCSP responder because there is no way for the certificate user to determine if the OCSP responder itself has been revoked. The actions needed to revoke an OCSP responder are similar to the actions needed to remove a trust anchor.

The real utility of OCSP lies in the single-response extension fields. If an application is checking a purchase order signature, the OCSP responder could provide a response stating that the certificate is not revoked and that the signature is acceptable for the stated dollar amount. CRLs cannot provide this additional context-specific functionality.

PKI MANAGEMENT PROTOCOLS

A CA needs to obtain the subscriber's public key, authenticate the subscriber's identity, verify that the subscriber possesses the corresponding private key, and verify any additional subscriber and key information before it signs a certificate. If the certificate contains incorrect information, a certificate user may establish security services with the wrong user or employ the public key for an inappropriate application. A CA must also determine that the status of a certificate has changed before it adds the certificate to the CRL. If the CA adds a valid certificate to the CRL, subscribers are denied service. If the CA fails to add a certificate whose status has changed to the CRL, certificate users will accept the invalid certificate. To meet these requirements, the CA must obtain trustworthy information from PKI participants. *PKI management protocols* are used by CAs to collect the information needed to issue certificates and CRLs. There are several PKI management protocols. Management protocols support two basic types of transactions: *certificate requests* and *revocation requests*.

As noted earlier, a CA needs to obtain trustworthy information before issuing or revoking a certificate. It may obtain this information from three PKI participants: the

prospective certificate holder, a current certificate holder, or a registration authority (RA). The CA has a different relationship with each of these participants.

A prospective certificate holder is essentially unknown to the CA but has requested acceptance into the PKI. The potential subscriber would like the CA to issue a certificate containing a specific identity and public key. The prospective certificate holder can provide this information in an initial certificate request, but the CA cannot determine from the data itself whether the name is appropriate. A CA can cryptographically verify that the requester has possession of the private key, however. For signature keys, the requester can simply digitally sign the request. For key management keys, a challenge-response mechanism may be required.

A certificate holder that possesses a currently valid certificate may request a new certificate. The requested certificate may have a different public key or include new name forms. The CA knows the subscriber's identity; otherwise, it would not have issued the current certificate. The current key pair may be used for authentication, and, as described previously, the CA can also cryptographically verify that the requester possesses the private key. The CA might not trust its subscribers to claim new names, however.

A certificate holder that possesses a currently valid certificate may also request revocation of one of his or her current certificates. The CA should always revoke a certificate upon the request of the certificate holder, so the signed request contains all the information required by the CA. This does not necessarily mean that the CA trusts the subscriber for this information or that the subscriber is telling the truth. If the holder of a private key asserts that it is no longer valid, this request must be honored. If the signed request came from another source, then the private key has been compromised, and the certificate must be revoked anyway.

The RA is empowered by the CA to collect information and verify its correctness. For certificate request operations, the RA may verify the prospective subscriber's identity and their e-mail address or other contact information. For revocation requests, the RA may identify the certificate subject and verify the reason for revocation. The RA is generally a certificate holder as well. RA digital signatures allow the CA to authenticate messages readily from the RA. An RA can review the documentation and determine whether a CA should honor a request.

PKI management transactions must be designed so that the CA obtains reliable transaction information. For some transactions, the CA and the certificate holder can implement the transaction without assistance. These are *two-party transaction models*. In other cases, the transactions leverage an RA to fill in the gaps in the trust relationships between the CA and prospective subscriber. These are *three-party transaction models*. The following is a brief survey of common PKI management protocols.

PKCS #10

Public Key Cryptography Standard (PKCS) #10, Certification Request Syntax Standard, describes a message syntax for certification requests. The certification request consists of a distinguished name (DN), the public key, an optional set of attributes, an algorithm identifier, and a digital signature. The optional attributes were designed to convey attributes for inclusion in the certificate (for example, an e-mail address), to provide the CA with additional information (for example, a postal address), and to establish a *challenge password* for use in a subsequent revocation request. The request is signed by the entity requesting certification using the corresponding private key. This signature is intended to achieve private key proof-of-possession.

PKCS #10 defines the syntax of a single request message, not a full protocol. The contents or format of the response is outside the scope of PKCS #10, although a PKCS #7 message is suggested as one possibility. Almost every PKCS #10 implementation employs PKCS #7 to return the certificate. The syntax and protocol used to request certificate revocation is also unspecified. PKCS #10 must be used with other message formats and protocols to provide functionality of a complete PKI management protocol.

PKCS #10 was not designed to be algorithm independent. The specification assumes the private key may be used to generate a digital signature, as is the case with the RSA algorithm. Proof-of-possession for key agreement algorithms, such as Diffie-Hellman, are outside the scope of PKCS #10. Proof-of-possession can be achieved using optional attributes to convey additional information, however. Despite these limitations, PKCS #10 remains the most widely used certificate request tool.

Certificate Management Protocol

When the IETF PKIX Working Group began development of a protocol for PKI management, they decided not to leverage PKCS #7 and PKCS #10. At the time, RSA Security held the copyright for the PKCS documents, so the IETF could not have change control. In addition, the working group wanted to develop a comprehensive protocol to support a broad variety of models, including RA participation, and implement algorithm-independent proof-of-possession. At the time, it was unclear whether PKCS #7 and #10 were an appropriate starting point to meet these goals.

The PKIX Working Group developed a new protocol defined by the combination of the Certificate Management Protocol (CMP; in RFC 2510), and the Certificate Request Management Framework (CRMF; in RFC 2511). The resulting protocol is comprehensive, can support practically any RA issuance model, and supports algorithm-independent proof-of-possession. The protocol also includes its own cryptographic message protection format, and it supports four transport protocols.

CMP defines seven transaction sequences, employing both request and response messages. These message pairs support three types of certificate requests, a CA certificate request, revocation, and key recovery operations. A proof-of-possession challenge sequence is defined for use in conjunction with the certificate request messages. The complexity of the CMP messages means that different implementations may not support the same combination

of optional fields. As a result, conforming implementations may not interoperate for all possible transactions. Nonetheless, the CMP specification defines the five common transactions in detail. These transactions are mandatory for conforming implementations, and they are specified in sufficient detail to achieve interoperability. Unfortunately, the *revocation request* and *revocation response* is not among these five transactions.

CMP messages are designed to handle multiple requests in a single message. This feature permits a user with two key pairs (one for signature and another for key management) to submit a single request. This feature also permits batch processing by RAs.

Certificate Management Messages over CMS

Over time, the IETF PKIX Working Group grew and became more diverse. Not everyone was comfortable with the direction of the CMP protocol. A group emerged that felt that it was crucial to leverage the installed base of PKCS #7 and PKCS #10. To these vendors, CMP represented a radical departure from a working, deployed protocol. CMP defined too many messages, and the CMP transactions demanded too many roundtrips. In their eyes, the comparative complexity of CMP overwhelmed the new functionality.

The PKIX Working Group fragmented into two camps. Those with a significant investment in CMP pointed out the weaknesses in PKCS #7 and PKCS #10, as well as the historic intellectual property issues. Those with a significant investment in PKCS #7 and PKCS #10 pointed out that a majority of PKIs used the RSA algorithm exclusively and that most PKIs did not involve RAs in protocols directly. When RSA Security decided to relinquish change control for PKCS #7 and PKCS #10 to the IETF S/MIME Working Group, RSA Security also resolved the intellectual property issues.

Eventually, a truce was achieved, and a second protocol emerged. Both protocols share the same certificate request format. The second protocol would use the new Cryptographic Message Syntax (CMS; in RFC 2630) specification to provide cryptographic protection for messages. The Certificate Management Messages over CMS (CMC; in RFC 2797) references PKCS #10 for a basic certificate request format, and CRMF for the more fully featured certificate request format. Furthermore, CMC offers algorithm independence and include support for direct involvement of RAs.

CMC specifies only two complete transactions: the simple enrollment protocol and the full enrollment protocol. These transactions each require two messages. The CA determines which type of certificate request has been received from the content itself. CMC began as a simple structure, but complexity was added to provide all of the necessary security and functionality. CMC control attributes determine the overall control flow. CMC defines 24 control attributes. These control attributes provide many of the features found in the CMP header, such as nonces and transaction identifiers. They are also used to implement proof-of-possession, pass arbitrary data, and indicate which extensions should appear in a certificate.

Simple Certificate Enrollment Protocol

Cisco Systems developed the Simple Certificate Enrollment Protocol (SCEP), and, as such, it is not an open standard like CMP or CMC. SCEP supports the secure issuance of certificates to network devices in a scalable manner, and it makes use of existing technology whenever possible. The existing technology includes the RSA algorithm, the DES algorithm, the PKCS #7 and PKCS #10 message formats, hypertext transfer protocol (HTTP), and LDAP. The protocol supports four transactions: distribution of CA and RA public keys, certificate requests, certificate queries, and CRL queries. The latter two transactions are actually repository functions, but they are included in the SCEP specification. The protocol also supports out-of-band revocation requests by establishing a revocation challenge password during the certificate request.

SCEP requires that end systems obtain three pieces of information as an initial condition: the CA IP (Internet protocol) address or domain name, the CA HTTP script path, and the URL for CRL queries if CRLs will be obtained from a directory. The end system uses an unprotected HTTP Get operation to obtain CA and RA certificates. At this point, the end system must contact the CA operator through out-of-band means and verify the hash of the certificate to ensure the integrity of this operation.

End entities begin the PKI enrollment process by generating their own public/private key pair, then they issue themselves a self-signed certificate. This certificate will be used for both authentication and key management, so the RSA algorithm is used. This provides each entity with a temporary, syntactically correct X.509 certificate. This step is required, because a digital signature protects all messages in the certificate request transaction, or signed and encrypted using PKCS #7, which assumes that a public key certificate is available. PKCS #7 requires an issuer name and serial number to identify the certificate.

POLICIES AND PROCEDURES

The bulk of this chapter focuses on PKI technical mechanisms, but technical mechanisms are insufficient on their own; they must be used in combination with a set of procedures to implement a particular security policy. Two documents describe the policies and procedures associated with a PKI. The first document is known as a *certificate policy* (CP), and the second is called a *certification practices statement* (CPS). These documents share a common format but have different audiences and different goals. RFC 2527, the *Certificate Policy and Certification Practices Framework,* established the recognized format for both a CP and a CPS.

Most PKI users will not refer to the CP or CPS. Users usually obtain the policy information indirectly by processing the certificate policies, policy mapping, and policy constraints extensions. There is a direct relationship between the contents of these extensions and the CP and CPS, however.

The CP is a high-level document that describes a security policy for issuing certificates and maintaining certificate status information. This security policy describes the operation of the CA, as well as the users' responsibilities

for the requesting, using, and handling of certificates and keys. The CP asserts that this security policy shall be implemented from certificate generation until its expiration or revocation. It does not specify how the policy shall be implemented. For example, a CP might state the following: "All subscribers shall be authenticated in person by an RA before a certificate is issued." The CP excludes all operational details, because these may evolve over time. The CP should not identify the physical location of the CA or the products used in the CA. By excluding these details, the CP is a stable and high-level document. Multiple CAs may operate under a single CP. This is often the case when multiple CAs are maintained by a single enterprise, jointly supporting a single community.

Different people will use the CP for different reasons. For example, the CP will be used to guide the development of the CPS for each CA that operates under its provisions. CAs from other enterprise PKIs will review the CP before cross-certification. Auditors and accreditors will use the CP as the basis for their review of CA operations. Application owners will review a CP to determine whether these certificates are appropriate for their application.

The CPS is a highly detailed document that describes how a particular CA implements a specific CP. The CPS identifies the CP and specifies the mechanisms and procedures that are used to achieve the security policy. The CPS asserts that the specified products will be used in combination with the specified procedures. The CPS might state the following: "Users will receive their certificates and smartcards from the RA after presenting the following credentials in person: (a) current driver's license, (b) work identification card, (c) blood sample, and (d) hair sample." A CPS includes sufficient operational details to demonstrate that the CP can be satisfied by this combination of mechanisms and procedures.

Each CPS applies to a single CA. The CPS may be considered the overall operations manual for the CA. Specific portions of the CPS may be extracted to form the *CA Operator's Guide, RA Manual, PKI Users Guide,* or other role-specific documentation. Auditors and accreditors will use the CPS to supplement the CP during their review of CA operations. Note that a CPS does not need to be published. The combination of a CP and the results of an accreditation process should be sufficient for external parties.

RFC 2527 proposes an outline with eight major sections and 185 second- and third-level topics. RFC 2527 established an outline with the following major sections:

Introduction
General Provisions
Identification and Authentication
Operational Requirements
Physical, Procedural, and Personnel Security Controls
Technical Security Controls
Certificate and CRL Profiles
Specification Administration
Privilege Management

Organizations seek improved access control. Public key certificates can be used to authenticate the identity of a user, and this identity can be used as an input to access control decision functions. In many contexts, however, the identity is not the criterion used for access control decisions. The access control decision may depend on role, security clearance, group membership, or ability to pay.

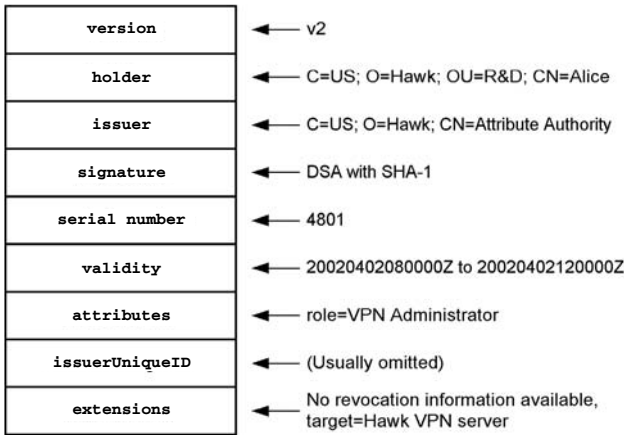


Figure 6: X.509 attribute certificate structure.

Authorization information often has a shorter lifetime than the binding of the subject identity and the public key. Authorization information could be placed in a public key certificate extension; however, this is not usually a good strategy. First, the certificate is likely to be revoked because the authorization information needs to be updated. Revoking and reissuing the public key certificate with updated authorization information can be expensive. Second, the CA that issues public key certificates is not likely to be authoritative for the authorization information. This results in additional steps for the CA to contact the authoritative authorization information source.

The X.509 *attribute certificate* (AC) binds attributes to an *AC holder*. Because the AC does not contain a public key, the AC is used in conjunction with a public key certificate. An access control function may make use of the attributes in an AC, but it is not a replacement for authentication. The public key certificate must first be used to perform authentication, then the AC is used to associate attributes with the authenticated identity.

ACs may also be used in the context of a data origin authentication service and a non-repudiation service. In these contexts, the attributes contained in the AC provide additional information about the signing entity. This information can be used to make sure that the entity is authorized to sign the data. This kind of checking depends either on the context in which the data is exchanged or on the data that has been digitally signed.

Figure 6 illustrates an attribute certificate for Alice. This is a version 2 AC, and the AC holder is Alice. The AC was issued by the Hawk Data *Attribute Authority,* and was signed with DSA and SHA-1. The serial number is 4801, and the AC is valid from 8 a.m. on April 2, 2002, until noon that same day. The attributes indicate that Alice is VPN administrator. The AC extensions indicate that this certificate is targeted toward the Hawk VPN server, and that revocation information is not available for this certificate. ACs often have no revocation information.

ACs may be short- or long-lived. In Figure 6, the AC permits Alice to administer the VPN for 4 hours. As a result of the short validity period, the AC issuer does not need to maintain revocation information. By the time revocation information could be compiled and distributed, the AC would expire. So, with short-lived ACs, revocation information is not distributed. If an AC has a longer life span (for example, weeks or months), then the organizations would need to maintain AC status information.

An AC can be obtained in two ways. The AC holder may provide the AC; this is known as the *push model*. Alternatively, the AC is requested from the AC issuer or a repository; this is known as the *pull model*. A major benefit of the pull model is that it can be implemented without changes to the client or to the communications protocol. The pull model is especially well suited to interdomain communication.

The AC is linked to a public key certificate in one of two ways. The AC holder can contain the issuer and serial number of a particular public key certificate, or the AC holder can contain a subject name. In the first case, the AC is linked to a specific public key certificate. In the second case, the AC is linked to a particular subject, and the AC may be used in conjunction with any public key certificate held by that subject.

FUTURE DEVELOPMENTS

One of the criticisms of PKI is that CRLs can become too large. When this happens, the overhead associated with CRL distribution is unacceptable. *Sliding window delta CRLs* can be used to reduce this overhead. Another criticism of PKI is that certification path construction and validation can be difficult. By delegating these functions to a trusted server, the amount of processing an application needs to perform before it can accept a certificate can be significantly reduced. Sliding window delta CRLs and *delegated path validation* are not widely deployed today, but they are likely to be employed in the future.

Sliding Window Delta CRLs

For PKIs that rely on CRLs, the challenge is to provide the freshest information to certificate users while minimizing network bandwidth consumption. Unfortunately, when PKIs rely on full CRLs, these requirements are in direct conflict. To maximize the freshness, CRLs must be updated frequently. As the time interval between updates shrinks, the probability that a client will find a useful CRL in its cache diminishes. At the extreme, certificate users will download a full CRL for each certificate validation. Most of the information on the CRL is the same, and identical information is transmitted repeatedly, consuming bandwidth without providing any benefit. To minimize the consumption of network bandwidth, CRLs should have reasonably long lifetimes. As the time interval between updates grows, the greater the probability that relying parties will have the appropriate CRL in their cache.

In the simple case, *delta CRLs* and full CRLs are issued together, and the delta CRL lists all the certificates revoked since the last full CRL was issued. A certificate user, who has the previous full CRL, may obtain complete information by obtaining the delta CRL and combining it with the already cached, previous full CRL. The certificate user obtains the freshest information available but consumes a fraction of the bandwidth. If the certificate user does not have the previous full CRL, the full CRL must be downloaded.

A sliding window delta CRL lists all the certificates revoked since an earlier full CRL, perhaps six generations earlier. This delta CRL may be combined with any of the full CRLs from the previous six generations. By repeating some of the revocation information in the delta CRL, there is a greater likelihood that the certificate user will have an acceptable full CRL in the cache, yet the amount of repeated information is small enough to avoid consuming significant bandwidth.

Most of the PKI-enabled applications do not exceed the limitations of full CRLs. As a result, delta CRLs are not widely deployed. Few commercial PKI client implementations process delta CRLs. Fewer CA products can generate sliding window deltas. As PKIs grow, however, the incentive to deploy innovative certificate status will likely grow.

Delegated Path Validation

Some PKI implementers want to offload the entire certification path construction and validation process to a trusted server. A relying party would provide a validation server with an end-entity certificate, one or more trust points, and the initial values for certification path validation, then the path validation server would respond with a message informing the relying party whether the certificate was acceptable. Standard protocols for these services have not yet been developed. This work is currently underway in the IETF PKIX Working Group.

Delegating the certificate validation process to a trusted server has a number of advantages. The certificate user achieves path construction and validation with a single roundtrip protocol, and then the certificate user verifies a single digital signature on the response. The single roundtrip is especially important in bandwidth-limited environments, especially wireless environments. If the certificate user has limited processing power, the reduction in signature verifications is also significant.

Delegating the certificate validation process to a trusted server may also provide performance advantages. If the path validation server has cached the necessary certificates and CRLs, the path validation server may be able to construct and validate a certification path quickly.

These benefits are not free. The path validation server performs all of the security-relevant operations. The path validation server must be secure, because it is the sole trust point for the relying party. In addition, some of the performance enhancements are based on the ability of the server to obtain and cache information. PKIs that rely on OCSP may be counterproductive to this model. In such a case, the path validation server is not likely to hold the required status information. The server will have to retrieve revocation information from the OCSP responder for each certificate in the certification path, mitigating much of the performance gain.

Performance is not the only reason to centralize certification path validation. Some organizations want impose a centralized management discipline with consistent policy enforcement. If applications use the same trusted path validation server, consistent results across the organization are ensured.

GLOSSARY

Attribute authority An entity that is responsible for the issuance of attribute certificates, assigning privileges to the certificate holder.

Attribute certificate A data structure that is digitally signed by an AA that binds attribute values with identification about its holder.

Certificate policy A named set of rules that indicates the applicability of a certificate to a particular community or class of application with common security requirements.

Certificate revocation list (CRL) A digitally signed list of certificate serial numbers associated with a set of certificates that are no longer considered valid by the certificate issuer.

Certification authority An entity that is responsible for the issuance of public key certificates, trusted by one or more certificate users.

Certification practices statement A description of the practices followed by a certification authority in issuing and managing public key certificates.

Public key certificate A data structure that contains a user identity, the user's public key, and other information, digitally signed by the CA.

Online certificate status protocol (OCSP) response A digitally signed response from a trusted server that implements the OCSP that provides status information for a queried certificate.

CROSS REFERENCES

See *Digital Signatures and Electronic Signatures; Electronic Payment; Guidelines for a Comprehensive Security System.*

FURTHER READING

Adams, C., & Farrell, S. (1999). *Internet X.509 public key infrastructure—Certificate management protocols* (RFC 2510). Retrieved March 2, 2003, from http://www.ietf.org/rfc/rfc2510.txt

Adams, C., & Lloyd, S. (1999). *Understanding public-key infrastructure.* Indianapolis, IN: Macmillan.

Chokhani, S., & Ford W. (1999). *Internet X.509 public key infrastructure—Certificate policy and certification practices framework* (RFC 2527). Retrieved March 2, 2003 from http://www.ietf.org/rfc/rfc2527.txt

Cooper, D. (2000, May). An efficient use of delta CRLs. *Proceedings of the 2000 IEEE Symposium on Security and Privacy* (pp. 190–202), Los Alamitos, CA: IEEE Computer Society Press.

Housley, R. (2002). *Cryptographic message syntax (CMS)* (RFC 3369). Retrieved March 2, 2003, from http://www.ietf.org/rfc/rfc3369.txt

Housley, R., & Polk, T. (2001). *Planning for PKI.* New York: Wiley.

Housley, R., Polk, W., Ford, W., & Solo, D. (2002). *Internet X.509 public key infrastructure—Certificate and certificate revocation list (CRL) profile* (RFC 3280). Retrieved March 2, 2003, from http://www.ietf.org/rfc/rfc3280.txt

International Telecommunication Union-Telecommunication Standardization Sector (ITU-T). (2000). *The directory—Authentication framework* (ITU-T Recommendation X.509).

Kaliski, B. (1998). *PKCS #7: Cryptographic message syntax, version 1.5* (RFC 2315). Retrieved March 2, 2003, from http://www.ietf.org/rfc/rfc2315.txt

Kaliski, B. (1998). *PKCS #10: Certification request syntax, version 1.5* (RFC 2314). Retrieved March 2, 2003, from http://www.ietf.org/rfc/rfc2314.txt

Liu, X., Madson, C., McGrew, D., & Nourse, A. (2001, September 11). *Cisco Systems' simple certificate enrollment protocol (SCEP)* (work in progress). Retrieved March 2, 2003, from http://www.vpnc.org/draft-nourse-scep

Myers, M., Adams, C., Solo, D., & Kemp, D. (1999). *Internet X.509 certificate request message format* (RFC 2511). Retrieved March 2, 2003, from http://www.ietf.org/rfc/rfc2511.txt

Myers, M., Ankney, R., Malpani, A., Galperin, S., & Adams, C. (1999). *X.509 Internet public key infrastructure—Online certificate status protocol (OCSP)* (RFC 2560). Retrieved July 30, 2002, from http://www.ietf.org/rfc/rfc2560.txt

Myers, M., Liu, X., Schaad, J., & Weinstein, J. (2000). *Certificate management messages over CMS* (RFC 2797). Retrieved from March 2, 2003, http://www.ietf.org/rfc/rfc2797.txt

Public Networks

Dale R. Thompson, *University of Arkansas*
Amy W. Apon, *University of Arkansas*

INTRODUCTION

Networks for the transfer of data between computers, both public and private, are ubiquitous in today's business world. A public network is one that is publicly available to subscribers (Stallings, 2001). It provides service to multiple subscribers and is built and maintained by a public network provider. Internationally, the term "public network" is often applied to networks that are under government control or are a national monopoly. However, a network can also be a privately owned network whose services are sold to the public. Whether the network is under government control or is a privately owned network whose services are sold to the public, businesses access the network by installing an access device at each site and using an access line to the nearest point of presence (POP) of the public network provider (Panko, 2001).

This chapter gives an overview of public network concepts and services and describes the structure of the public switched telephone network (PSTN) system, the technologies used both for access to a public network and within the public network itself, issues related to choosing a public or a private network, and public networks in the Internet and e-commerce environments.

OVERVIEW OF PUBLIC NETWORK CONCEPTS AND SERVICES

Traditionally, companies desiring to connect business computers in different geographic locations have used private networks. That is, they have used point-to-point leased lines between business sites to create their own circuit-switching or packet-switching networks for their data communication requirements (Panko, 2001). Unlike telephone calls, which set up the required capacity as needed, leased lines provide dedicated transmission capacity between sites. These networks are called private networks (Stallings, 2001). By using leased lines, companies have a network capacity that is always available and are offered volume discounts for the bandwidth available on the leased line. An example of a private network is shown in Figure 1.

There are several disadvantages to private networks. Private networks require higher initial costs. The leased line connections must be planned and installed. The switching devices must be provided. And, once a network is operational there are ongoing management and maintenance costs of the networks (Panko, 2001). A public network is an alternative to a private network.

There are advantages to using a public network. A public network does not require a complex network of leased lines and switching devices that the business must plan and install. There is commonly one access line installed per site. Even if a leased line is used to connect to the nearest POP, there are usually less leased lines required. For example, if there are 10 sites using the public network, then there are 10 leased lines. Compare this to a fully meshed private network that requires 45 leased lines. For N locations, $N(N-1)/2$ leased lines are required for a connection to and from each site. Even if not every site is connected to every other site in the private network, but sites are connected through intermediate sites, the number of leased lines for a public versus a private network is generally smaller. Finally, because of competitive pricing, public networks are less expensive than private networks (Stallings, 2001). Figure 2 illustrates an example of a public network.

The global Internet is a network that is publicly accessible worldwide. The Internet is not one single network, but is composed of several networks connected together and communicating with standard Internet technologies (Moody, 2001). Access to the Internet is achieved via an Internet service provider (ISP). The Internet allows a business to have a worldwide presence. Through the use of

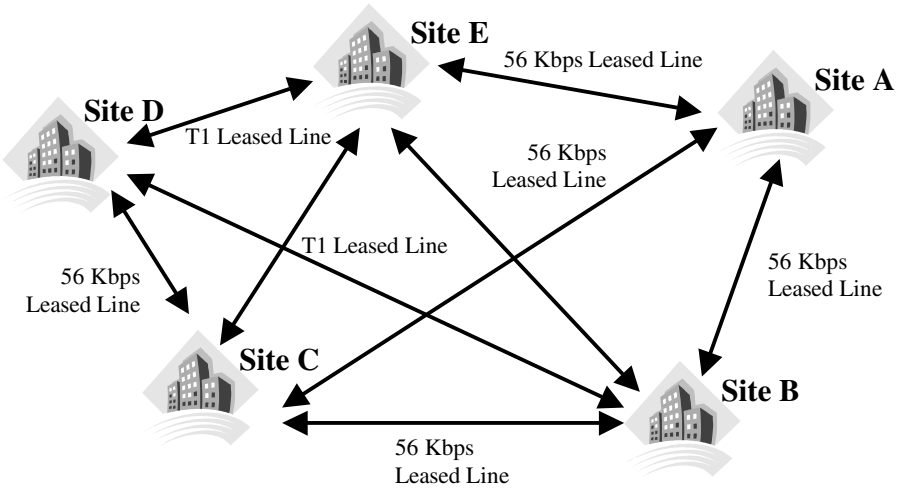


Figure 1: A private switched data network.

E-commerce purchases can be made automatically with software.

A network that transfers data and information only within a single business is called an intranet (Moody, 2001). Intranets use the same technologies as the Internet but access is restricted to employees. They carry corporate information that can range from being routine such as e-mail, manuals, and directories or can be sensitive information such as that of project management and internal purchasing. An intranet can be built using a private or a public network. A private network is naturally an intranet. A business using a public network can ask that the data be restricted to only go to other locations of the same business. Of course, the bandwidth is still being shared with other businesses that use the same public network.

An extranet is a hybrid between the public Internet and the private intranet (Moody, 2001). A portion of the intranet is extended to business partners in a controlled and restricted way. The extranet can be used for project management of projects between partners. Another common and practical use of the extranet is to allow partners access to the stock levels and shipping status. Direct online purchasing of supplies and other applications are made possible through the use of an extranet.

The global Internet can be used to provide an intranet or an extranet by creating a virtual private network (VPN). A VPN is a private network that is deployed over public facilities, but provides the same levels of privacy, security, quality of service, and manageability as private networks (Cisco, 2001).

A VPN can be created when all sites are already connected to the Internet. With a VPN, hosts at different sites communicate across the Internet using either a tunnel mode between local networks, or by using a direct transport communication. However, there are two serious problems that can occur with VPNs since the company no longer has control of the entire data network (Panko, 2001). One problem is the security of the data, because the Internet was not designed to support secure transmission. This problem can be solved through the use of encryption and by using tunnel mode for communication. A second problem is congestion on the Internet. Congestion can

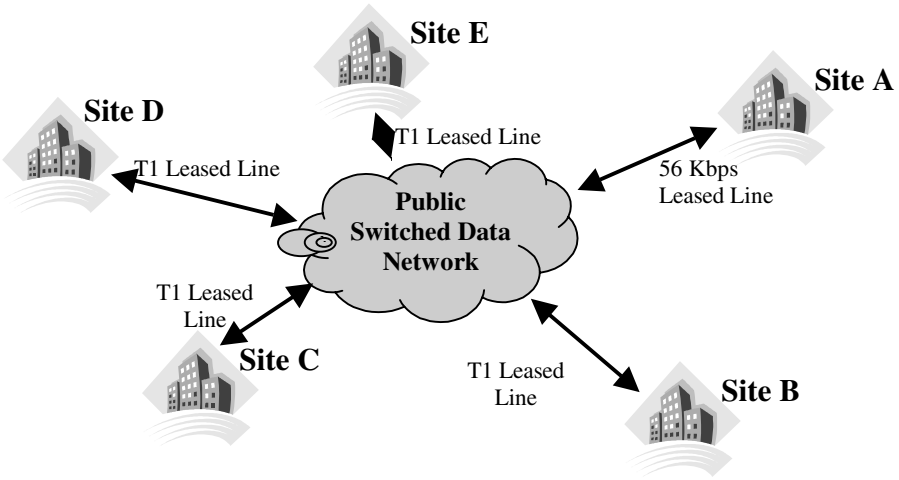


Figure 2: A public switched data network.

cause data to be delayed or even lost. A VPN uses a public network for site-to-site communication and added technology to solve the problems of security and congestion (Panko, 2001).

A public network provider has a value-added network if it owns the packet-switching nodes and leases transmission capacity from an interexchange carrier such as AT&T (Stallings, 2001). It is called a value-added network because the leased lines add value to the packet switching nodes. A network provider that provides a value-added network is sometimes called a value-added carrier. In many cases a public network provider will partner with companies that provide services that require network connectivity such as Web hosting and give discounts to them for using their network. A business which bundles a service with a particular public network provider is called a value-added reseller.

Public network providers often offer services such as Web hosting to subscribers in addition to connectivity between sites. These additional services are called value-added services. These services include asset management, configuration control, fault management, monitoring, Web-based reporting, Web hosting, e-mail services, and content delivery networks.

Asset management is keeping inventory of devices that are connected to the network. As devices are added or taken off the network the asset management system will keep an up-to-date log of the assets. Configuration control is about maintaining and keeping records of the configuration of networked devices. The network provider typically maintains the configuration of the packet switching node that connects each of the subscriber locations to the network. A provider will also monitor devices to detect faults and either fix them or notify the appropriate on-site personnel. This is called fault management. A provider can invest in large network operation centers for monitoring their subscribers' network devices. This includes maintaining a firewall to prevent unwanted users into the network and intrusion detection systems for detecting activity that is consistent with common hacker techniques. With Web-based reporting the provider gives the subscriber reports about the status of their network and a history of its downtime and performance.

One of the most popular value-added services is Web hosting. The provider maintains one or more servers and allocates space on them for the subscriber's Web site. The provider maintains the server and performs backups. Subscribers are given access to their portions of the server to post their Web sites and control their content. An advantage to using this value-added service is that it is likely that the subscriber has other sites that are connected to the same public network. If the server is connected to the same public network, it provides faster response times to the end users.

Medium to large users who have high volumes of content serving a distributed set of users may consider a value-added service called a content delivery network (CDN). A CDN intelligently distributes the content to multiple locations and closer to the end user. By moving the customized content closer to the end user the end user receives faster response times (Allen, 2001). Queries to the main server or group of servers are routed to the location that can best respond to the query. Content is cached at each of the locations and future requests are serviced more quickly because the information traverses fewer links in the network. There are three main advantages to a CDN. First, end users receive faster response times. Second, it relieves congestion on the original server that maintains the master copy of the content. Finally, it reduces the amount of data transmission capacity required on the network since the content is distributed to multiple locations and does not have to come from the original server. Some of the popular CDN providers are Akamai (http://www.akamai.com) and Mirror Image (http://www.mirror-image.com).

STRUCTURE OF THE PUBLIC SWITCHED TELEPHONE NETWORK SYSTEM

The public switched telephone network system is often used to provide the technology that a business uses to access a public network or is the technology of the public or private lines. The structure of the PSTN in the U.S. has evolved from one that was almost entirely controlled by a single company to one that allows competition in a free market. Before January 1, 1984, AT&T (also known as the Bell System) controlled 80% of the PSTN in the U.S. (Bellamy, 2000). A Justice Department antitrust suit filed in 1974 and a private antitrust case by MCI resulted in a breakup of AT&T (Noam, 2001). The suit argued that AT&T used its control of the local operation as an unfair advantage against competing long distance carriers.

On January 1, 1984, AT&T was divided into smaller companies. The breakup involved the divestiture of seven Bell operating companies (BOCs) from AT&T. The seven regional BOCs were known as "Baby Bells" or regional BOCs (RBOCs) and initially carried only regional telephone and mobile service. The network was partitioned into two levels (Bellamy, 2000), and the remaining part of AT&T retained the transport of long distance telephone service.

The U.S. was divided into local access and transport areas (LATAs), which are controlled by local exchange carriers (LECs). LECs can transport telephone calls within a LATA, also called intra-LATA traffic, but are not permitted to transport traffic between different LATAs, also called inter-LATA traffic, even though the same BOC may control both LATAs. The inter-LATA traffic is transported by interexchange carriers (IXCs), commonly known as long distance carriers. Each IXC interfaces at a single point in the LATA called a point of presence. At divestiture, AT&T became an IXC and it opened the door to competition for other companies' long distance service. The major IXCs in the U.S. include AT&T, MCI–WorldCom, and Sprint.

The divestiture decree was supervised by District Judge Harold Greene and known as the modified final judgment (Noam, 2001). LECs had to grant equal access to all IXCs. The service offered by the LECs to the IXCs had to be equal in type, quality, and price (Bellamy, 2000). Also, users could specify their "primary" IXC to transport their long distance and international calls (Noam, 2001). Or,

users could use other IXCs on a call-by-call basis by dialing a prefix.

Another major change in the U.S. PSTN occurred with the 1996 Telecommunications Act that amended the Communications Act of 1934 (Noam, 2001). RBOCs had to comply with a list of tasks before they were permitted to provide long-distance service within their regions. The list permitted competition in the RBOCs regions. It was argued that it was necessary to induce competition in these local markets. RBOCs were required to provide interconnection to new market competitors, unbundle their network, permit competitors to resell their service, and provide users with number portability.

The new local service providers became known as competitive local exchange companies (CLECs) (pronounced "see-lecks") (Noam, 2001). The incumbent LECs became known as ILECs. For a CLEC to be competitive with the ILEC requires that it be able to interconnect with the users cost effectively. Therefore, there came a great struggle between CLECs and ILECs on the issue of collocation since the ILEC had a significant advantage with the existing network. In "physical collocation" a CLEC places its cables and equipment inside the ILEC's central office (CO) to hand off traffic. In another arrangement called "virtual collocation" the physical handoff of the traffic occurs inside or outside the CO, but uses ILEC-owned equipment and must be the economic equivalent of "physical collocation."

It may appear from the previous discussion that the breaking up of the U.S. PSTN is relevant only to the United States but the trend is happening in other parts of the world as well (Noam, 2001). Japan opened its markets to competition. Also, the Europeans have privatized their service. Noam argues that at first a network is not feasible unless supported by outside sources such as governments. As the network grows the average costs decline initially and then rise as a few high-cost users are added. Without regulation the network would not grow beyond a certain point because of the high cost of adding these high-cost users. From a political and societal point of view the network becomes a necessity instead of a convenience and should be offered to everyone. Therefore, the monopolistic breakdown of the network is caused by its own success.

ACCESS AND PUBLIC NETWORK TECHNOLOGIES

To use a public network for data services, a user must access the public network through some network service from the user's computing equipment to the nearest public network node. Factors in selecting a particular service include the cost of the service that is provided and the features, including the transmission speed, that are provided by the technology. Generally, the higher the transmission speed that a technology can support, the more costly the service becomes. Transmission speeds for networks are described in bits per second. Unlike when memory size is described, 1 Kbps is exactly equal to 10^3 bits per second, 1 Mbps is exactly equal to 10^6 bits per second, and 1 Gbps is exactly equal to 10^9 bits per second.

Many technologies are available for access to a public network and for use within the public network. The most inexpensive network access is through a voice-grade modem. A modem is used to convert a digital computer signal to an analog signal that can be sent across ordinary telephone lines. Voice-grade modems can receive data at up to 56 Kbps. In contrast, digital lines that are used to access the network range in transmission speed from 56 Kbps to 10 Gbps. Within the public network a few technologies, including X.25, frame relay, asynchronous transfer mode (ATM), and synchronous optical network (SONET), have become the most commonly used technologies. Table 1 lists the most common technologies along with a comment about usage. Table 1 also compares the transmission speed and the time to download a 10-megabit (1.2 Megabyte) file.

Voice-Grade Modems

A modem is the most inexpensive and easiest to use access technology. The use of modems for data transmission will be substantial for many years to come (Stallings, 2001). Voice-grade modems use a 4-KHz bandwidth on an ordinary telephone line, the same bandwidth that is used for voice signals. Modems can be packaged inside an information product, such as a personal computer. Companies often have modem banks that allow employees to dial-in directly to the company intranet or to access a large computer system.

On March 1, 1993, the International Telecommunications Union (ITU) Telecommunications Standardization Sector (ITU-T) was created as a permanent organ of the ITU, an agency of the United Nations. The charter of the ITU-T is to standardize techniques and operations in telecommunications. Several standard specifications for voice-grade modems have been designated by the ITU-T. Two of the most significant modem specifications are V.32, which is a dial-up modem that transmits at 9600 bps, and V.90, also a dial-up modem. V.90 sends at 33.6 Kbps and receives at 56 Kbps, the highest rates available for voice-grade modems (Stallings, 2001).

Digital Subscriber Lines

A faster service than voice-grade modems that is beginning to be offered by telephone companies is the digital subscriber line (DSL). A widely publicized version of this is asymmetric digital subscriber line (ADSL). ADSL offers high-speed downstream access to the customer site, and a lower speed upstream access from the customer. The ITU-T has developed a standard for low-speed ADSL called G.992.2, or G.Lite. G.Lite specifies downstream speeds of 1.5 Mbps, but sometimes lower downstream speeds are used. Most users find asymmetric speeds to be acceptable, since upstream traffic frequently consists of keystrokes or the transmission of short e-mail messages, whereas downstream traffic may include Web pages, or large amounts of data. In addition to data speed, an advantage of DSL over voice-grade modems is that DSL modems allow voice traffic to be multiplexed onto the telephone wires coming into the customer site. A customer can talk on the telephone at the same time that data are being transferred.

Table 1 Common Network Technologies

Service	Usage Comments	Transmission Speed	Download
Voice-Grade Modem	Modems are inexpensive, telephone rates reasonable for modest connect times	Upload: Up to 33.6 Kbps Download: Up to 56 Kbps	3 min or more
Digital Subscriber Line	More expensive than voice-grade modems, downlink rates higher than uplink	Upload: From 16 Kbps to 640 Kbps Download: From 768 Kbps to 9 Mbps	1.1–13 s
Cable Modems	Download rates depend on the number of simultaneous customers and configuration	Upload: From 64 Kbps to 256 Kbps Download: From 10 Mbps to 30 Mbps	0.3–1 s
Satellite	A cost-effective choice in remote locations	Upload: From 56 Kbps to 256 Kbps Download: From 150 Kbps to 1 Mbps	10–67 s
Integrated Services Digital Network	Charges generally based on duration of call	Basic rate: 128 Kbps, higher rates available	1.3 min
Digital leased lines: 56 Kbps (DS0), T1 (DS1), T3 (DS3), . . .	Most common leased line for high-traffic voice and data; fixed price for a specific capacity	DS0: 56 Kbps T1, DS1: 1.54 Mbps T3, DS3: 44.7 Mbps	56 Kbps: 3 min T1: 6.5 s T3: 0.22 s
SONET	Specification for optical links, highest speed	From 155.52 Mbps to 2.488 Gbps leased	0.004–0.06 s
X.25	Older technology, still in use in public networks	56 Kbps, but can be slower or faster	3 min or more
Frame Relay	Fixed price per month for a specific capacity, widely installed and used	From 16 Kbps to 44.736 Mbps	0.22–625 s
ATM	Universal technology for wide area networking	From 1.544 Mbps to 2.5 Gbps for access	0.004–6.5 s

The telephone company does not have to install any special equipment to use voice-grade modems. However, when the telephone company offers DSL service it has to install digital subscriber line access multiplexers at the end offices. Figure 3 illustrates the equipment used for DSL (Panko, 2001). Because special equipment has to be installed, DSL service is not available in all areas. One factor that determines the availability of ADSL is the distance to the central office. In general, if the distance is greater than 18,000 feet ADSL service is not available. Also, the prices are fluctuating as DSL becomes available in more and more areas.

Cable Modems

Cable modems are a service offered by cable television companies. Often, the cable television or telephone

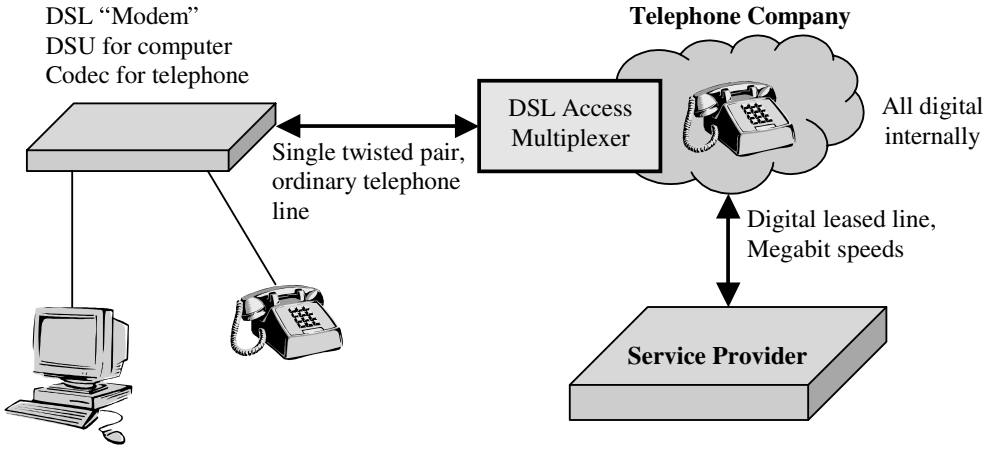


Figure 3: Asymmetric digital subscriber line. Source: *Buisness Data Communications and Networking, 3/E* (Panko, 2001). Reprinted by permission of Pearson Education Inc., Upper Saddle River, NJ.

company operates as both a transmission carrier and a network provider. As with ADSL, the downstream speed of cable modem is much faster than the upstream speed. The upstream speeds are similar to ADSL, but the downstream speeds can be several times faster. However, multiple customers on the same cable line share the capacity. When many customers are accessing the network at the same time the real downstream transmission speed can be much lower. If network traffic is bursty, though, the chances are unlikely that all customers are downloading at exactly the same moment so that sharing does not become as issue until about 100 customers share the same cable service (Panko, 2001).

Satellite

An often cost-effective alternative for network access is the use of satellite technology. This may be particularly true in areas where other wire-based technologies are not yet available. For example, many rural areas do not have the density of potential users that can justify the cost of installation of wire-based technologies such as DSL or cable modems.

Satellites are characterized by the type of orbit they use. The most common type of satellite is the geostationary satellite. These satellites orbit the Earth at about 22,300 miles directly above the equator at exactly the same speed as the Earth's rotation. Because of this, the satellite always appears to be in the same position in the sky and tracking of the satellite by stations on Earth is simplified (Stallings, 2001). The disadvantage of this type of satellite is that the propagation time it takes for the signal to be sent from a transmission station on the Earth to the satellite, and then to be received back on the Earth is about 0.24 s. For large data downloads this is not noticeable since the time overlaps with the time to receive the entire message. However, for interactive computer use or for applications such as telephone calls the time is noticeable and can be annoying. In addition, geostationary satellite signals are not received well in very far northern or southern regions of the Earth.

Two other types of orbits include low- and medium-Earth orbiting satellites. This technology is being proposed for use with mobile terminals and remote locations that need stronger signals and less propagation time. Successful businesses that use this technology are rare. One company currently operating under bankruptcy regulations, Iridium, provides global, mobile satellite voice and data solutions with complete coverage of the Earth through a constellation of 66 low-Earth orbiting satellites (Iridium, 2002).

Large satellite dishes create narrow footprints for transmission, and large dishes are used for point-to-point trunk transmissions. A small dish creates a very large footprint that is suitable for television broadcasts in a large region. Today, very small aperture terminal systems are available and provide a low-cost alternative to expensive point-to-point satellite connections. These stations share satellite transmission capacity for transmission to a hub station (Stallings, 2001).

Satellite access has some advantages over wire-based technologies. The technology is available now for all locations in the United States, whereas DSL and cable modem technologies may not be available in some locations for some time. For the speeds and services available the technology is cost-competitive. However, in order to use satellite, the user must have a clear view of the southern sky. The uploads speeds are modest, so satellite is not suitable for businesses that require high-upload bandwidth for applications such as large upload data transfers or for hosting Web sites. Also, the download bandwidth is shared with all users at the site, and so the technology is not currently suitable for more than five simultaneous users.

At least one company offers packages with two-way, always-on, high-speed Internet access via satellite that is specifically designed to meet the needs of small businesses (StarBand, 2002). StarBand uses a 24-by-36-inch dish and a special modem at the customer's site to connect the user's site to the network. StarBand also serves as a network provider. Fees include an initial equipment fee and a monthly fee for access. Value-added services such as domain registration and networking support for setting up small office networks can be a part of the package.

Integrated Services Digital Network

Many telephone companies offer integrated services digital network (ISDN), a digital service that runs over ordinary telephone lines. As with voice-grade modems the ITU-T has set standards for ISDN. ISDN can be used as an access technology and within a public network. Basic ISDN service includes two "B" channels, each at 64 Kbps, and a "D" channel that is used for signaling. It is possible to use one "B" channel for voice and one for data, but most service providers bond the two "B" channels together to provide a 128 Kbps data rate. Standards for higher rates also exist. Like ADSL, ISDN requires that the telephone company install special equipment at the end office before an ISDN service can be offered. A special ISDN "modem" is used at the customer site.

ISDN is the result of efforts in the early 1980s by the world's telephone companies to design and build a fully digital, circuit-switched telephone system (Tanenbaum, 1996). Because ISDN is circuit-switched, there is never any congestion on the line from the customer to the network service provider. However, since data traffic is generally bursty the user pays for bandwidth that may not be used. ISDN is expensive compared to the modest gain in transmission speed. The customer generally has to pay for the ISDN line to the telephone company and then has to pay an additional fee to a network service provider. The use of ISDN is likely to decline as other higher speed and more economical technologies become available.

Digital Leased Lines

In terms of number of circuits, the most common leased lines are 56 Kbps (Panko, 2001). The transmission capacity of a 56 Kbps is actually 64 Kbps but one bit out of eight is used for signaling, leaving the user with 56 Kbps. A 56 Kbps line is the same as digital signal zero (DS0). The next higher transmission speed is a T1 (DS1), which provides 1.544 Mbps. While a 56 Kbps leased line is relatively inexpensive, the difference in cost and performance between a 56 Kbps and a T1 line is large. Therefore, fractional T1's are also available at 128 Kbps, 256 Kbps, 384 Kbps, and so on. In Europe and other parts of the world

a different digital hierarchy of transmission capacities is used. The standards are defined in the Council of European Postal and Telecommunications authorities (CEPT). The E1 standard operates at 2.048 Mbps and is analogous to the T1 standard. The next step is a T3 (DS3) at 44.7 Mbps and the corresponding CEPT E3 standard operating at 34.4 Mbps. Higher transmission capacities are available using synchronous optical network (SONET) and the synchronous digital hierarchy (SDH) and range from 155.52 Mbps to 10 Gbps.

Digital leased lines can be used to build a company's leased line private network, as shown in Figure 1, or can be used in combination with a public network, as shown in Figure 2. When leased lines are used to access a public network the traffic between several sites must be multiplexed over the single access line. Therefore, it is important to be sure that the leased line is fast enough to support this traffic. For example, if a site has 15 56 Kbps leased lines connected point-to-point with other sites and wants to convert this to a single access line to a public network, then the access line would require at least 840 Kbps of capacity. From Table 1, this would require a T1 line (Panko, 2001).

Synchronous Optical Network

Synchronous optical network defines a hierarchy of standardized digital data rates. A compatible version, Synchronous digital hierarchy has been published by the ITU-T. SONET is intended to provide a specification for high-speed digital transmission over optical fiber.

SONET, or SDH, is the highest speed and most costly digital leased lines. SONET/SDH operates in multiples of 51.84 Mbps. Standards are specified as OCx for SONET, and STMx for the SDH specification. A common SONET/SDH speed is OC3/STM1, at 156 Mbps. Other common rates include 622 Mbps, 2.5 Gbps, and 10 Gbps. SONET technology can be used for access both to the public network and within the public network.

X.25

X.25 was developed during the 1970s for use in public packet switching networks, and this standard was later ratified by the ITU-T (Tanenbaum, 1996). X.25 was very slow, often running at only 9600 bps, but it was fast enough for the text-based transmissions of early networks. Its use is declining, but it is still popular in the U.S. for low-speed applications such as a department store's point-of-sale transaction network. Also, there are many X.25 legacy connections, particularly in Europe and in countries where the telecommunications infrastructure is lagging. X.25 is one of a few standards that have been set by the ITU-T for public switched data networks. Other standards set by the ITU-T for public networks include ISDN, frame relay, and ATM.

Frame Relay

Frame relay is the most popular technology choice within public switched data networks today (Panko, 2001). Its speed range matches the needs of the greatest corporate demand, and it has very competitive pricing. Frame relay can also be used instead of leased lines as an access technology or to connect company private networks. Its low overhead even makes it suitable for interconnecting LANs and high-speed stand-alone systems (Stallings, 2001). Current commercial offerings of frame relay include MCI–WorldCom, which offers frame relay service access speeds from 28.8 Kbps to 45 Mbps (MCI–WorldCom, 2002), and Qwest, which offers frame relay service access speeds from 64 Kbps to 45 Mbps (Qwest, 2002).

Typically, a company accesses a public frame relay network through a leased line. Several frame relay virtual circuits are multiplexed over a single access line to the public network. A virtual circuit is a connection from source to destination and represents an end-to-end path that all packets from the same source to the same destination go through. Virtual circuits simplify forwarding decisions and make the costs of the switches cheaper. A permanent virtual circuit (PVC) is one that is set up manually when a company first subscribes to a public network, and only changes when the site changes. For a large company network, a PVC is established for every pair of sites that would get a leased line in a private leased line network.

The frame relay protocol includes functions for detection of transmission errors and congestion control functions. The frame relay protocol allows users to negotiate a committed information rate (CIR) when a connection is set up. The CIR is the network's commitment to deliver data in the absence of errors, and represents the user's estimate of its "normal" traffic during a busy period. Any traffic sent above the CIR is not guaranteed to arrive, but may arrive if the network has the capacity to deliver it. In addition, a maximum allowable rate is defined, and all traffic above this level is discarded (Frame Relay Forum, 2002).

Pricing for frame relay is usually divided into several different components. First, the company needs a frame relay access device. This is a router that has been modified to allow it to communicate with the frame relay's first switch. Second, the company must lease an access line to the nearest POP of the public network. If the POP is a long distance away then the customer must use expensive, long-distance access lines. The leased line must be fast enough to handle the available bit rate on the line.

At the POP, the leased access line connects to a port on the frame relay switch of the public network. The fee for the port is usually the largest single element in frame relay pricing. To prevent wasting port capacity, the speed of the leased line should be at least as fast as the port speed. There is usually a monthly fee for each PVC and this fee depends on the speed of the PVC. Finally, some vendors build in other fees, such as per-bit traffic charges or fees to set up and tear down switched virtual circuits that are established on a call-by-call basis. Frequently there are substantial initial charges to install the access device, leased line, port connection, or PVC. Figure 4 illustrates the pricing elements in frame relay (Panko, 2001).

Asynchronous Transfer Mode

Asynchronous transfer mode is now viewed to be the universal technology for networking and will likely replace many other current offerings (Stallings, 2001). Just as frame relay allows messages to be divided into many

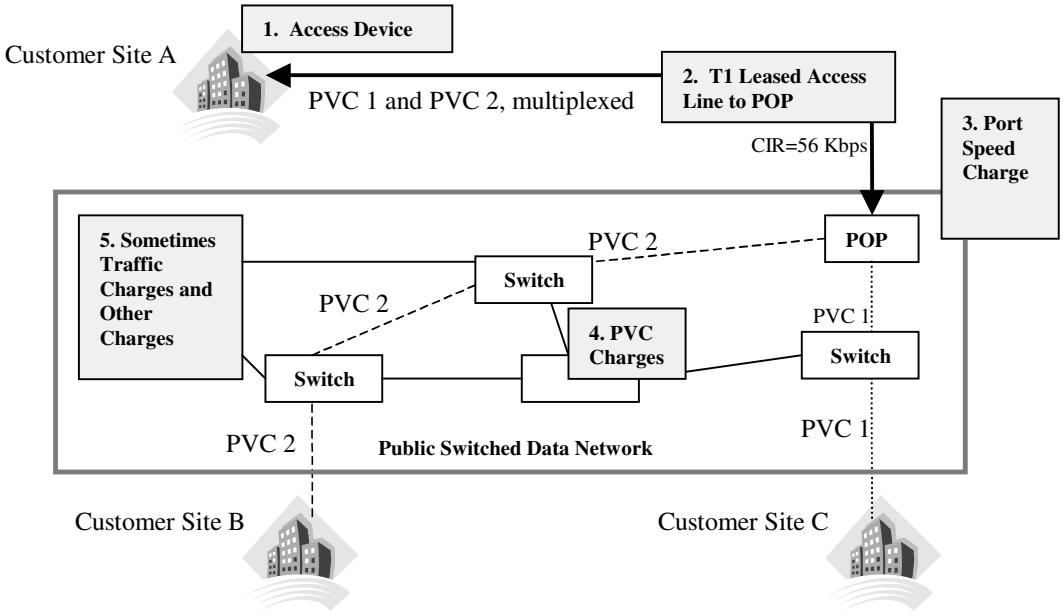


Figure 4: Pricing elements in frame relay services. Source: *Buisness Data Communications and Networking, 3/E* (Panko, 2001). Reprinted by permission of Pearson Education Inc., Upper Saddle River, NJ.

frames that can be sent across a switched network, ATM uses cell relay. Like frame relay, ATM multiplexes many logical connections over the same physical interface, sending information in fixed size 53-byte cells. ATM can support data, video, voice, and Internet traffic on a single access line.

The use of cells in ATM allows many important features to be defined for a virtual channel. For example, users can negotiate the ratio of cells lost to cells transmitted, cell delay variation and parameters such as the average rate, peak rate, burstiness, and peak duration for a virtual channel (ATM Forum, 2002). The ATM service can use permanent virtual channels for static connections. ATM also allows switched virtual channels to be set up dynamically on a call-by-call basis.

Four classes of ATM service have been defined (Stallings, 2001):

Constant bit rate: The network provider ensures that this rate is available, and the customer is monitored to be sure the rate is not exceeded.

Variable bit rate (VBR): A sustained rate for normal use is defined, and a faster burst rate for occasional use is also defined. The faster rate is guaranteed, but not continuously. The ATM Forum divides VBR into real-time VBR (rt-VBR) and nonreal-time VBR (nrt-VBR) (ATM Forum, 2002). With rt-VBR the application has tight constraints on delay and delay variation, but the rate is allowed to vary according to parameters specified by the user. The nrt-VBR is for applications that are bursty, but do not have tight constraints on delay and delay variation.

Available bit rate (ABR): The user has a guaranteed minimum capacity. When additional capacity is available on the network, the user may burst above this without risk of cell loss.

Unspecified bit rate (UBR): Cells are delivered with best effort, meaning that any cell may be lost. The main difference between UBR and ABR is that ABR provides feedback to the user so that the user can control the amount of data being sent and reduce the risk of loss.

ATM is a high-performance service and is expensive. In the range of speeds where ATM speeds overlap with frame relay, frame relay is more attractive because it is cheaper. However, as customer needs increase, ATM becomes a more attractive option. ATM is widely used within high-speed public networks and by companies that need higher speed private networks. Most ATM public switched data network providers currently offer speeds from 1 Mbps to 156 Mbps, with higher speeds coming. These public networks require access lines ranging from T1 to a SONET OC-3 line. MCI–WorldCom offers ATM access speeds from 1.544 Mbps to 622 Mbps (MCI–WorldCom, 2002). Qwest offers ATM access speeds from 1.544 Mbps to 155 Mbps (Qwest, 2002).

CHOOSING A PRIVATE NETWORK OR A PUBLIC NETWORK PROVIDER

There are several categories to consider when one decides whether to use a private network or a public network. If a public network is chosen, then these same categories can help in choosing a network provider. A survey ISPs conducted in 2001 found that the top three areas that differentiated the best ISPs from the rest were reliability, performance, and low cost (Greenfield, 2001). Subscribers to ISPs in the survey also considered support to be important. In addition, network control is a factor in deciding whether to choose a private network or a public network. Other factors mentioned in the survey include breadth of service, security, installation, repairs, and remote access.

Reliability

Reliability is defined as the amount of time the network service is available. Reliability can be difficult to evaluate because several different things can cause downtime. For example, if a user is trying to transfer data from a server that is down then from the user's point of view the network is down. When a packet switching node or dedicated leased line in a large complex network does fail it affects a large amount of transmission capacity and therefore a large number of users. For example, MCI–WorldCom's frame relay outage in August 1999 lasted eight days and affected 30% of MCI's frame relay customers, perhaps as many as 70,000 users (Orenstein and Ohlson, 1999).

An advantage to using a private network is that the redundancy of the network can be designed according to the business requirements. The major disadvantage is that it requires investment in redundant packet switching nodes and leased lines for fault tolerance, personnel training, disaster recover planning, and testing. These expenses are often overlooked or have less priority when a private network is designed (Snow, 2001). Or once the private network is operational these expenses are considered low priority. Therefore, when there is an outage the business is not prepared for it and its effects are worse than if a disaster recovery plan had been written.

The reliability of a public network has advantages and disadvantages. The advantage of using a public network is that since the cost is spread out over several subscribers added investment in reliability can be cost effective. The disadvantage is that a subscriber is completely dependent on the provider for reliable service. Service level agreements have to be negotiated with clear and strict penalties if the provider does not meet the negotiated reliability. If reliability is of high importance to a business, then they may subscribe to two or more public network providers for added reliability.

Cost and Performance Tradeoffs

The choice between a public and private network includes determining the tradeoffs between the cost and performance of the network. The performance of the network is defined by throughput and delay. The throughput is the actual data speed seen by the user in bits per second. The delay is the maximum end-to-end delay that a packet will incur in the network.

The costs of the network may vary depending on the type and volume of traffic that the network will carry. The type of traffic on a network is classified as being either stream or bursty (Stallings, 2001). Stream traffic is long and relatively constant and therefore more predictable than bursty traffic. An example of stream traffic would be voice traffic or uncompressed video. Bursty traffic is short and sporadic such as computer-to-computer communication in the Internet. Although sporadic, bursty traffic often requires a large transmission capacity for brief periods of time. Many Internet applications such as the Web and e-mail create such bursty traffic. If there are several bursty traffic sources that share a communications link and the volume of the combined traffic is high then the aggregate traffic on the link may be considered stream traffic.

Bursty traffic requires a different type of network than stream traffic. For example, if one file is required to be transferred from an office to a central site once a day then a dial-up connection may be the most feasible. On the other hand, if there is bursty traffic to be transferred among a small number of sites and the aggregate of the bursty sources has a high volume then a private packet switching network would be more efficient. Leased lines are not dependent on volume but have a constant fixed rate for a given transmission capacity and distance. If the percentage of use of the leased line is high enough then the volume discount given by the constant fixed rate can be cost effective. For example, large nationwide contracts can negotiate T1 access lines for $200 a month while users in metropolitan areas can get T1 access for approximately $900 per month (The Yankee Group, 2001). Compare this to $50 per phone time's 24 channels that is $1,200 per month for an equivalent amount of bandwidth.

If there is a moderate volume of bursty traffic to be transferred among a medium to large number of sites then a public network may be a better choice. Since the public network provider has several subscribers, the aggregate volume of traffic is great enough to have high use and therefore is cost effective for the provider. These savings are passed on to subscribers who do not have enough volume of traffic to justify a private network.

The costs for some network technologies can be negotiated with the expected performance in mind. For example, with frame relay, the user chooses the committed information rate in bits per second and committed burst size (Frame Relay Forum, 2002). A frame relay network provider will also specify a maximum end-to-end delay for a frame in their network. These parameters are a part of the pricing for frame relay service.

The price of a network is usually divided up into a fixed cost and a variable cost. The fixed access cost depends on the type of access technology that a user connects to the POP with and the distance the user is from the POP. There may not be a variable cost, but if there is the price is dependent on the volume of traffic. A user may subscribe to a certain data rate from the network for a fixed cost and if the user exceeds the limit, the user is charged for the additional usage.

Support

Support is defined as the quality of a provider's technical and logistical help. In one survey the complaint most cited was the lack of support (Greenfield, 2001). Networks are complex and they do break and fail. A good network provider should be fast to respond and correct problems. A business should consider where the nearest technician would be coming from to service their sites. Service level agreements will define minor and major problems and the type of responses that the network provider will provide.

Control

An organization relies on its network to operate its business (Stallings, 2001). Management requires control of the network to provide efficient and effective service to the organization. There are tradeoffs between a private and public network when considering control. There are

three areas of control that need to be considered: strategic control, growth control, and day-to-day operations.

Strategic control of a network is designing and implementing a network to satisfy the organization's unique requirements. If the organization operates its own private network then it can determine the configuration of the network. But, if the organization uses a public network the organization does not have strategic control over the configuration of the network. The public network provider designs the network for the average subscriber.

Growth control of the network is the ability to expand and make modifications to meet the changing requirements of the organization. It includes adding switching nodes and leased lines, modifying the capacities of the leased lines, and changing the network technology. A private network provides the maximum flexibility for growth control since the organization has complete control over the network. If an organization is a subscriber to a public network it has almost no growth control. All requirements are constrained by the capabilities of the public network.

The other type of control is the day-to-day operation of the network. This includes the ability to handle traffic during peak times, to diagnose problems, and to repair problems quickly. In a private network the organization sets the priorities of the day-to-day operation to fit their business. But, with a private network they also have to hire or develop in-house expertise to maintain the often complex network. Also the organization has to address the reliability of the network by determining where to install redundant packet switching nodes and dedicated leased lines. If an organization is a subscriber to a public network then it is dependent on the public network provider. There are peak traffic times and the public network provider may focus its efforts on the overall health of the network and not on an individual user. On the other hand, the provider can afford more redundancy and hire or develop more in-house expertise because these costs are spread out over several subscribers.

Other Factors

Other factors that are important in choosing a network solution include breadth of service, security, installation, repairs, and remote access. Many network providers offer a wide breadth of value-added services, as previously described. A provider that can provide value-added services such as Web hosting bundled with its network service can have a big advantage. If the server is on the same network that other customers are connected to then performance is better.

Security of a network includes restricting access to information located on corporate servers and preventing malicious activities like denial-of-service attacks that shut down a Web site. A network provider can provide firewalls to restrict activity to sites, VPNs to encrypt and restrict access between sites, and intrusion detection to detect malicious activity.

The installation and repairs category includes the timeliness and quality of an installation. Networks are complex and often require coordination between multiple organizations. For example, in the U.S. if a leased line crosses two different LATAs then at least one local provider and at least one IXC will be required. Also, realistic time schedules are important because a rushed installation usually results in a poor quality installation and long-term problems.

For many businesses remote access is important to be competitive. Remote access permits users in a business to communicate often with e-mail and to access corporate data. Remote access is dependent on the number and location of the network provider's in-dial modem pools. If this is an important part of the business model then a business should look for a provider that has multiple access points in the areas that their employees travel.

PUBLIC NETWORKS IN THE INTERNET AND E-COMMERCE ENVIRONMENTS

Public networks provide a cost-effective solution for small businesses to connect to the Internet and participate in E-commerce because they provide connections to the public Internet through one or more locations. Access to the Internet is restructuring the marketing, sales, management, production, accounting, and personnel management in businesses (Moody, 2001). The Internet provides online up-to-the-minute reports for marketing. Marketing can easily monitor their competitors by accessing the online information and competitors can easily monitor a business. The Internet has had two effects on sales. First, a business can have a worldwide presence. Second, customers are demanding the correct information for deciding which business to buy from. The online purchase is now being handled automatically by software (e-commerce). Members of the sales department can access corporate information over the network while on the road. Management can now have access to more of the organization. They can access information from marketing, sales, production, accounting, and personnel including previous years' sales and regional performance of a product. They can have online meetings and stay in contact with e-mail. Production can receive quicker feedback from the field and have feedback from suppliers about their stock levels. Accounting can pay online and receive up-to-the-minute information. Personnel information such as directories can be provided online and manuals and training material can be placed online.

CONCLUSION

Public networks are an increasingly popular solution for businesses to link multiple sites together to exchange information and to connect to the Internet. Public networks offer several advantages over private networks composed of leased lines, including lower cost for a given performance, value-added services, and fewer requirements of maintaining in-house expertise for network maintenance, support, and similar administrative and management tasks. Public networks do have some disadvantages, including potential variation in performance due to congestion on the public network, and lack of control over day-to-day operations, upgrades, and long-range planning for capacity changes. However, public networks combine

connectivity with value-added services such as Web hosting and CDNs and are a good choice for many businesses.

In the future, only organizations with special requirements in the areas of performance, control, and security will continue to maintain and install private networks. Many organizations with private networks today will migrate their private networks to public networks or use VPNs via their Internet connection. Even organizations that continue to have private networks will have at least one connection to the one global public network called the Internet to participate in activities such as e-mail and E-commerce.

GLOSSARY

Asynchronous transfer mode A network technology, characterized by sending data in fixed size 53-byte cells and offering various levels of service.

Asynchronous digital subscriber line A digital service that uses ordinary telephone lines to connect a customer to a public network. Asynchronous DSL has download speeds that are much faster than the upload speeds.

Content delivery network (CDN) A value-added service that distributes the content to multiple locations and closer to the end user. By sophisticated caching schemes a CDN reduces response times.

Frame relay The most popular technology choice within public switched data networks. Data are divided into frames that are sent on switched networks.

Interexchange carrier A long-distance carrier in the public switched telephone network system.

Internet service provider An organization that provides access to the Internet by providing an Internet address and support of Internet protocols to the subscriber.

Leased line A digital line that provides dedicated transmission capacity between sites.

Local exchange carrier A carrier that controls traffic within a single local access and transport area.

Public network A network that is publicly available to subscribers. A public network can be under government control, operate as a national monopoly, or can be a privately owned network whose services are sold to the public.

Private network A business network composed of point-to-point leased lines between sites.

Public switched telephone network The network that makes up the public telephone system.

Value-added carrier A network provider that provides a value-added network.

Value-added network A network constructed by a network provider that owns the packet-switching nodes and leases transmission capacity to add value to the network.

Value-added reseller A business that provides a service (e.g., Web hosting) that requires network connectivity and sells it for use with a particular public network provider. The network provider often gives discounts to the business for using the network.

Virtual private network A network that uses a collection of technologies applied to the public network to provide the same levels of privacy, security, quality of service, and manageability as private networks.

CROSS REFERENCES

See *Integrated Services Digital Network (ISDN): Narrowband and Broadband Services and Applications; Virtual Private Networks: Internet Protocol (IP) Based; Wide Area and Metropolitan Area Networks*.

REFERENCES

ATM Forum (2002). Retrieved July 17, 2002, from http://www.atmforum.com

Allen, D. (2001, December 5). Content delivery networks come home. *Network Magazine*. Retrieved May 9, 2002, from http://www.networkmagazine.com/article/NMG20011203S0017

Bellamy, J. C. (2000). *Digital telephony* (3rd ed.). New York: Wiley.

Cisco (2001). Secure business communications over public networks. Retrieved April 4, 2002, from http://www.cisco.com/warp/public/cc/pd/rt/800/prodlit/sbcp_wp.htm

Frame Relay Forum (2002). Retrieved May 7, 2002, from http://www.frforum.com

Greenfield, D. (2001, September 5). Slugfest results. *Network Magazine*. Retrieved May 7, 2002, from http://www.networkmagazine.com/article/NMG20010823S0012

Iridium Satellite (2002). Retrieved May 7, 2002, from http://www.iridium.com

MCI–WorldCom (2002). Retrieved May 7, 2002, from http://www.worldcom.com

Moody, G. (2001). The business potential of the Internet. Retrieved December 12, 2001, from http://www.worldcom.com/generation_d/whitepapers/

Noam, E. M. (2001). *Interconnecting the network of networks.* Cambridge, MA: The MIT Press.

Orenstein, C. S., & Ohlson, K. (1999, August 13). MCI network outage hits Chicago trading board hard. *Computerworld*.

Panko, R. R. (2001). *Business data communications and networking.* New Jersey: Prentice Hall.

Qwest (2002). Retrieved May 7, 2002, from http://www.qwest.com

Snow, A.P. (2001). Network reliability: the concurrent challenges of innovation, competition, and complexity. *IEEE Transactions on Reliability, 50*(1), 38–40.

Stallings, W. (2001). *Business data communications.* New Jersey: Prentice Hall.

StarBand Communications (2002). Retrieved May 7, 2002, from http://www.starband.com

Tanenbaum, A. S. (1996). *Computer networks*. New Jersey: Prentice Hall.

The Yankee Group (2001, December 31). Endless pressure—Price and availability review for private lines and dedicated access services. Retrieved April 23, 2002, from http://www.yankeegroup.com

R

Radio Frequency and Wireless Communications

Okechukwu C. Ugweje, *The University of Akron*

INTRODUCTION

Radio-frequency (RF) wireless communication systems have been around for many years with applications ranging from garage door openers to satellite communication. The technology has been advancing at an unprecedented rate and its impact is evident in our daily lives. In many parts of the world, wireless communication is the fastest growing area of the communication industry, providing a valuable supplement and alternative to existing wired networks (Cellular Communications Services in the USA, 2003). Based on the number of subscribers to wireless communication products and services, it is now the preferred method of communication (Wireless Communications, Market & Opportunities, 2003). Many systems formerly carried over the wire are now carried over wireless media.

The remarkable success of cellular mobile radio technology has fundamentally changed the way people communicate and conduct business. The wireless revolution has led to a new multi-billion-dollar wireless communications industry. Linking service areas, wireless communication has altered the way business is conducted. For example, with a laptop computer, a wireless modem, and a cellular phone, a business consultant can contact his or her office and clients and conduct business while traveling. While traveling, field service and sales personnel can access corporate databases to check inventory status, prepare up-to-the-minute price and delivery quotes, modify schedule activities, and fulfill orders directly to the factory. Company personnel can use two-way paging services to stay in close contact, even when traditional wired communication services are available. Handheld hybrid phone-computer-fax machines feed information to wireless communication networks, allowing an executive to make decisions while on a leisure outing.

In this chapter, we present a concise summary of the subject of RF and wireless communication. This includes a discussion of the general concepts and definitions of RF wireless communication, various forms and applications of RF wireless communication, and the concepts, properties, and behavior of radio waves. We also summarize existing and emerging technologies for wireless communication. Of particular interest is the cellular mobile radio system, which has become the most widespread RF wireless communication system.

OVERVIEW OF RF WIRELESS COMMUNICATION

Introduction

Wireless or RF communication began at the turn of the 20th century, over 100 years ago, when Marconi

established the first successful and practical radio system. His experiment in 1895 demonstrated the transmission of radio signals a distance of 2 kilometers (Proakis & Salehi, 2002). He conducted additional experiments leading to 1901 when his radiotelegraph system transmitted radio signals across the Atlantic Ocean, from England to Newfoundland, about 1,700 miles away (Mobile Telephone History, 2002). However, only telegraphic codes were transmitted. On December 24, 1906, Reginald Fessenden accomplished the first radio communication of human speech over a distance of 11 miles from Brant Rock, Massachusetts, to ships in the Atlantic Ocean (Mobile Telephone History, 2002). Radio was no longer limited to telegraph codes; it was no longer just a wireless telegraph. This was a remarkable milestone highlighting the beginning of the voice-transmitted age.

In the early years of RF wireless communication, radio broadcasting was the most deployed wireless communication technology. The invention of the vacuum tube and vacuum triode hastened the advancement in radio transmission of voice signals. Radio broadcast by way of amplitude modulation and later frequency modulation (FM) was made possible. Amplitude modulation of the radio frequency was used to carry information until FM was introduced in the late 1930s (Mark & Zhuang, 2003). After FM was introduced, many other RF wireless systems such as television, one- and two-way radio, and radar were introduced between the late 1920s and the mid-1950s. Another milestone was witnessed in the late 1970s, which marked the beginning of the growth in cellular mobile radios and personal communication services. The first successful commercial analog cellular mobile telephone was demonstrated in 1979 (Durgin, 2003). Currently, wireless communication of all kinds abounds in our society.

System Architecture

In RF wireless communication systems, radio waves are used to transfer information between a transmitter (Tx) and a receiver (Rx). RF systems can be classified as either terrestrial-based or space-based systems. Terrestrial-based systems include microwave point-to-point, wireless local area networks, and cellular mobile radio, just to mention a few. Terrestrial microwave systems are limited in distance and line-of-sight (LOS) propagation may be required. Relay towers using carefully aligned directional antennas are often used to provide an unobstructed path over an extended distance. The data signal is processed, up- or down-converted, modulated or demodulated, filtered, and amplified at the transceivers. The transmitted signal propagates through the air and is attenuated by several propagation mechanisms discussed below.

Space-based systems (e.g., the satellite) are similar to terrestrial microwave systems except that signals travel from earth-based ground stations to a satellite (uplink) and a signal is sent back from the satellite to another earth-based ground station (downlink). This achieves a far wider coverage area than the earth-based systems. The satellite system could be in geostationary earth orbit, medium earth orbit, or low earth orbit.

A typical wireless communication system is shown in Figure 1. It consists of a source of information, a hardware subsystem called the transmitter, the channel or means by which the signal travels, another hardware subsystem called the receiver, and a destination of the information (the sink).

The source supplies the information to the transmitter in the form of audio, video, data, or a combination of the three. The Tx and Rx combination is used to convert the signal into a form suitable for transmission and

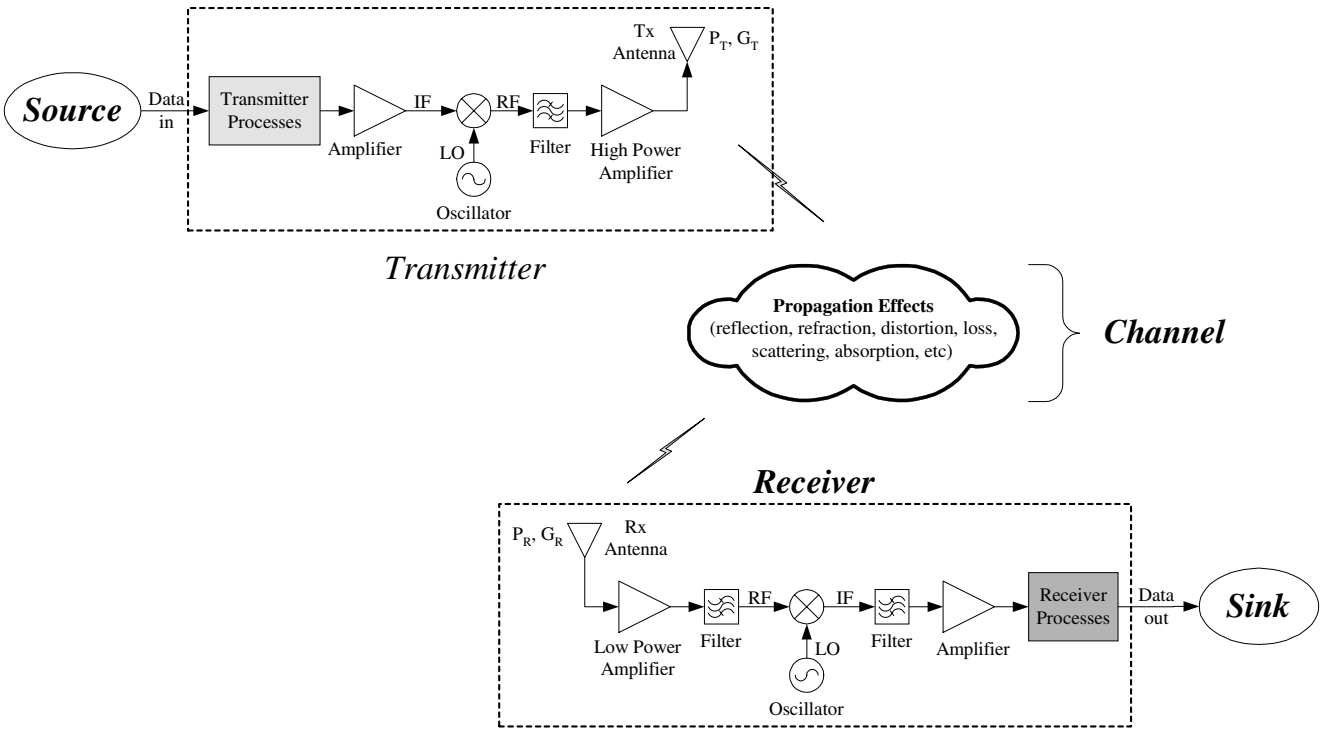


Figure 1: Simplified model of terrestrial-based RF wireless communication systems.

then to convert the signal back to its original form. This is achieved through the process of modulation (or encoding) at the Tx side and demodulation (or decoding) at the Rx side. The channel is the medium by which the signal propagates, such as free space, unshielded twisted pair, coaxial cable, or fiber-optic cable. In wireless communication the channel is free space. Noise and interference is added to the signal in the channel, which causes attenuation, distortion, and eventually error in the received signal.

The transmitter and receiver are very complex systems consisting of many internal components. A block diagram representation of some of the components is shown in Figure 1. Components are denoted as transmitter processes, receiver processes, amplifiers, mixers, local oscillators (LO), filters, and antennas. The transmitter processes represents functions of the transmitter such as modulation, encoding, analog-to-digital conversion, multiplexing, addressing, and routing information. The receiver processes, on the other hand, denote inverse functions such as demodulation, decoding, digital-to-analog conversion, and demultiplexing, as well as addressing and routing information. Effective transmission and reception of radio waves involves processes such as amplification and filtering of the signal at various internal stages, mixing of the desired signal with a local oscillator signal, translating the signal from one frequency to another, and transmission or reception of the RF energy through the antenna. The amplifier is characterized by its gain, noise figure (or output power), and linearity (Weisman, 2003). The gain (in dB) of the amplifier is a measure of how much bigger the output signal is than the input signal. The noise figure (or noise ratio) is a measure of the quality of the receiver system. Mixers are commonly found in the Tx and Rx subsystems and are used to create new frequencies or translate existing frequencies to new ones. They are sometimes called up or down converters. The most common translation of frequency is from intermediate frequency (IF) to RF and vice versa. The mixer performs this function by effectively multiplying two signals at two different frequencies. A signal source that provides one of the inputs to the mixer is the LO. A common type of LO is a voltage-controlled oscillator. The function of the filter is frequency selectivity. Filters select signals based on their frequency components. Regardless of the construction, all filters can be classified as lowpass, highpass, bandpass, or bandstop. These names are descriptive of the function of the filter. For example, a lowpass filter will select signals with low frequency and reject signals with high frequency. A special type of filter commonly used in RF systems is the duplexer. It is used to combine the functions of two filters into one. The duplexer facilitates the use of one antenna for both transmission and reception. The sink or destination can vary as much as the source and type of information.

In the channel, external noise in the form of manmade noise (generated by electrical manmade objects), atmospheric noise, and extraterrestrial noise is introduced. Atmospheric noise is produced by electrical activities of the atmosphere. This type of noise is predominant in the range 0–30 MHz and is inversely proportional to its frequency. Extraterrestrial noise is produced by activities of the cosmos, including the sun.

Radio Spectrum Classification

Radio frequencies or radio waves constitute the portion of the electromagnetic spectrum extending from 30 kHz to 300 GHz. The entire RF spectrum is classified into different bands and ranges, based on propagation properties. Baseband signals or source signals (e.g., audio signals) are in the low-frequency range below 30 kHz. This range of frequencies is classified as very low frequency (VLF), which must be translated into RF before transmission.

Radio waves are also described by their wavelength, λ, as belonging to a particular wavelength range such as shortwave, medium-wave, or millimeter-wave. The higher the frequency, the lower the wavelength, because $\lambda = c/f_c$, where $c = 3.9 \times 10^8$ m/s is the speed of light, and f_c is the carrier frequency. The wavelength is related to the realizable antenna length, L, system bandwidth, B, and other practical system parameters. For example, higher frequency radio waves produce smaller λ, require shorter L, have higher bandwidth efficiency, ρ, are more susceptible to fading, and suffer from atmospheric distortion. The characteristics and applications of radio frequencies are summarized in Table 1.

Within each frequency range, several bands of frequencies can be designated for communication. These bands are commonly identified by either f_c or a letter symbol, as illustrated in Figure 2 (Acosta, 1999; Federal Communications Commission, 1997). For example, in practical applications one could describe an RF system as operating in the C, X, K, or K_A band instead of using the actual frequency numbers. A complete list of the radio-frequency allocation can be found in *Selected U.S. Radio Frequency Allocations and Applications* (2002).

Because of the congestion or unavailability of usable spectrum at the lower frequency bands (below 20 GHz) and the recent demand for multimedia communication at high data-rate capabilities, system designers have directed their attention toward the use of SHF and EHF for communication (Acosta, 1999). Currently, there is a great deal of research on developing RF systems operating at frequencies above 20 GHz (K_A band and above) (National Aeronautics and Space Administration, 1998).

This interest in the EHF band is justified due to its potential benefits, such as the availability of usable spectrum, high data-rate capability, reduced interference, and high achievable gain with narrow beam widths of small antennas (Ippolito, 1989). The drawback, though, is that at these frequencies atmospheric distortion, especially rain attenuation, is very severe (Acosta & Horton, 1998; Xu, Rappaport, Boyle, & Schaffner, 2000). The severity of the meteorological effects increases with increasing frequency. At some frequency bands, the meteorological effects can cause a reduction in signal amplitude, depolarization of the radio wave, and increase in thermal noise (Ippolito, 1989).

Radio Wave Characteristics

When electrical energy in the form of high-frequency voltage or current is applied to an antenna, it is converted to electromagnetic (EM) waves or radio-frequency energy. At the Tx, the antenna converts a time-varying voltage or current into a time-varying propagating EM wave. The resulting EM wave propagates in space away from the

Table 1 Radio-Frequency Band Classification and Characteristics

Frequency Band	Frequency Range	Propagation Characteristics	λ	ρ	L	Typical Use
Very low frequency (VLF)	< 30 kHz	Low attenuation day and night; high atmospheric noise level.	Long	Low	Long	Baseband signals; power line; home control systems; navigation and submarine communication.
Low frequency (LF)	30–300 kHz	Slightly less reliable than VLF; absorption in daytime.				Long-range navigation; marine communication; radio beacons.
Medium frequency (MF)	0.3–3 MHz	Attenuation low at night, high in day; atmospheric noise.				Maritime radio; direction finding; AM broadcasting.
High frequency (HF)	3.0–30 MHz	Omni-directional energy radiation; quality varies with time of day, season, and frequency.				International broadcasting, military communication; long-distance aircraft and shipcommunication.
Very high frequency (VHF)	30–300 MHz	Direct and ground waves; cosmic noise; antenna design and height is critical.				VHF TV; FM broadcast; two-way radio, AM aircraft communication and navigational aids.
Ultra high frequency (UHF)	0.3–3 GHz	LOS; repeaters are used to cover greater distances; cosmic noise.				UHF TV; cellular phone; radar; microwave links; PCS.
Super high frequency (SHF)	3.0–30 GHz	LOS; atmospheric attenuation due to rain (>10 GHz), oxygen and water vapor.				Satellite and radar communication; terrestrial microwave; wireless local loop.
Extremely high frequency (EHF)	30–300 GHz	LOS; millimeter wave; atmospheric attenuation due to rain, oxygen and water vapor.	Short	High	Short	Experimental; wireless local loop.

source (the antenna) at the speed of light with the succeeding wave front changing in amplitude as the voltage or current changes in amplitude. Radio waves propagate through space as traveling EM fields proportional to the time-varying voltage or current. The propagating RF energy is composed of an electric field and a magnetic field component. The two fields exist together because a change in the electric field generates a corresponding change in the magnetic field, and vice versa. At the Rx the antenna performs an inverse operation of converting a time-varying propagating EM wave to a time-varying voltage or current.

Polarization of the radio wave is important and is given by the direction of the electric field component. Usually the construction and orientation of the antenna determine the electric field component. Many antennas are linearly polarized, either horizontally or vertically. The magnitude of the power radiated in the direction of propagation can be calculated as the effective isotropic radiated power (EIRP) or effective radiated power. This is the maximum radiated power available from a Tx in the direction of maximum gain for isotropic or directional antennas, respectively. It is a measure of the effectiveness of an antenna in directing the transmitter power in a particular direction (Rappaport, 2002).

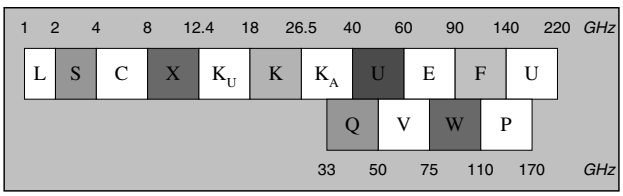


Figure 2: Typical symbol assignment for RF bands.

Forms of Radio Waves

Radio waves propagate in space in various forms. The characteristics of the propagating waves are of interest in many wireless communication systems designs. Propagating radio waves can be classified as direct (or free space), ground (or surface), tropospheric, and ionospheric. These types of waves are illustrated in Figure 3.

Direct waves are the simplest kind of radio waves, in which propagation is in free space without any obstruction. They are projected in a straight LOS between the Tx and Rx. The two-way radio, cellular mobile telephone, and personal communication system seldom have this type of radio wave.

Ground waves are confined to the lower atmosphere or the surface of the earth. A ground wave includes that

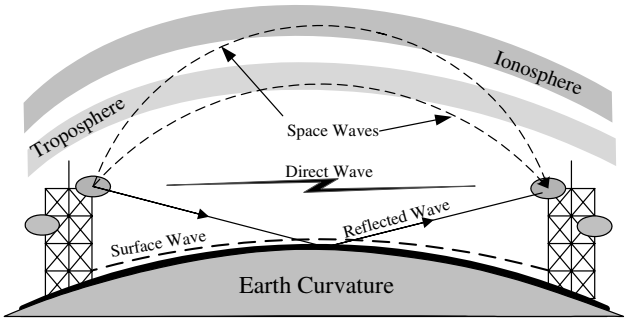


Figure 3: Common types of radio waves in wireless communication systems.

portion of the radio wave directly affected by terrain and objects on the terrain. It is guided along the surface of the earth, reflecting and scattering off buildings, vegetation, hills, mountains, and other irregularities on the earth's surface. These waves propagate outward from the antenna but undergo refraction due to variation in the density of the atmosphere (Garg & Wilkes, 1996). The signal strength decreases as the distance between the Tx and the Rx increases. This wave affects all frequencies in the MF, HF, and VHF ranges, and it is the dominant wave in cellular mobile radio systems. Vertical polarization, the direction of the electric-field component, is best for this type of wave. The polarization is determined by the construction and orientation of the antenna.

Tropospheric and ionospheric waves are commonly referred to as sky waves. They propagate in outer space but can return to earth by reflection or scattering either in the troposphere or in the ionosphere. The tropospheric wave is that portion of the radio wave close to the earth's surface as a result of gradual bending in the lower atmosphere (Garg & Wilkes, 1996). The bending action is due to the changing effective dielectric constant of the atmosphere through which the wave is passing. Its reflective index gradually decreases with height, resulting in a bending path taken by the wave. The troposphere extends about 10 miles above the surface of the earth and applies to waves with wavelength shorter than 10 m; i.e., $\lambda < 10$ m. The ionospheric wave is similar to the tropospheric wave except that it travels farther and the reflection occurs in the ionosphere, 40–400 miles above the earth. This wave is highly reliable for telemetry, tracking, weather forecasting, and tactical military applications. Note that different wavelengths are reflected to dissimilar extents in the troposphere and ionosphere.

Radio-Frequency-Based Systems

Figure 4 shows the different forms of RF-based wireless communication systems, which we have classified into six groups: microwave RF systems, fixed and mobile satellite systems, wireless networks and protocols, personal communication systems, remote sensing systems, and emerging technologies. No distinction is made between the communication layers and protocols in this classification. These systems transmit and receive radio waves tuned to specific bands of frequencies. Microwave is loosely used to describe all radio frequencies between 1 and 40 GHz. This includes the UHF, SHF, and EHF systems. The lower microwave frequencies, i.e., UHF, are most often used for terrestrial-based RF systems, whereas the higher microwave frequencies, i.e., SHF and EHF, are used for satellite communications. A terrestrial microwave system transmits carefully focused beams of radio waves from a transmitting antenna to a receiving antenna. A terrestrial microwave system uses LOS propagation to communicate between the Tx and the Rx with a typical distance of 30 miles between relay towers.

Personal communication services (PCS) are a new generation of wireless-phone technology that introduces a wide range of features and services greater than those available in analog and digital cellular phone systems (International Engineering Consortium, 2003a). It includes any system that provides people with access to information services, such as cellular telephones, home-based systems (cordless telephones, remote control, short-range two-way radio), beepers, pagers, and much more (Goodman, 1997; Rappaport, 2002). PCS provides the user with an all-in-one wireless phone, paging, messaging, and data

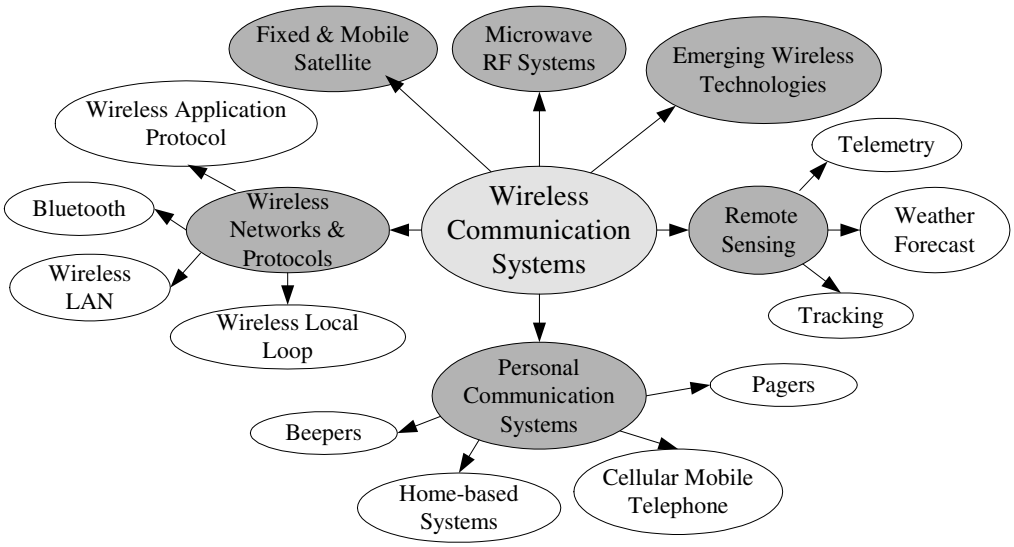


Figure 4: Different forms of RF-based wireless communication systems.

Table 2 Variants of Wireless LAN Systems and Bluetooth

Properties	IEEE 802.11	HiperLAN	Ricochet	HomeRF	Bluetooth
Spectrum (GHz)	2.400–2.4835; 5.15–5.35, 5.525–5.825	5.15, 17.1	0.902–0.928	2.404–2.478	2.402–2.480
Range	150 feet	150 feet	1000 feet	<150 feet	10cm–100 m
Power Consumption	Not specified	Not specified	Not specified	100 mW	1 mW, 10 mW and 100 mW
Energy Conservation	Directory based	Yes	Unknown	Directory based	Yes
Physical Layer	DSSS/ FHSS/IR	DFS with BPSK/QPSK/ QAM	FHSS 162 hops/s	FHSS 50 hops/s	FHSS 1600 hops/s
Channel Access	CSMA/CA	TDMA/TDD	TDMA	Hybrid TDMA and CSMA/CA	FHSS, Master slave TDMA
Mobility Support	Not specified	Yes	Yes	No	No
Raw Data Rate	2, 11, 6–54 Mbps	23.5, 54 Mbps	288 kbps	1 and 2 Mbps	1 Mbps
Traffic	Data (DCF)	Data	Data	Voice + Data	Voice or Data
Speech Coding	Unknown	OFDM	Not available	ADPCM, 32 bps	64 kbps with CSVD/log PCM
Security	40 bit RC4	DES, 3-DWS	RSARC-4	Blowfish	Minimal in PHY
Communication Technology	Peer-to-peer, MS-BS	Peer-to-peer, MS-BS	Peer-to-peer	Peer-to-peer, MS-BS	Master/slave

service. The most significant segment of this technology is the cellular mobile radio. It is the fastest growing segment of the telecommunications industry. Based on the number of new subscribers worldwide and the number of services, the cellular mobile radio system has evolved as the dominant wireless communication system. Its history dates back many decades, but the modern-day mobile radio became widespread in the 1980s (Rappaport, 2002). The cellular mobile radio system is discussed further below.

Wireless networks and protocols include systems such as wireless local area networks (W-LAN), wireless local loops (WLL), wireless application protocol (WAP), and Bluetooth. These systems are used mainly to provide data communication. W-LAN is an extension to, or an alternative for, a wired LAN. W-LAN provides the functionality of wired LAN, without the physical constraints of the wire itself, combining data connectivity with user mobility (Bing, 2000; Geier, 1999; Wenig, 1996). W-LANs have the potential to support user mobility and constant and unlimited access to information by linking several wireless devices to the wired infrastructure network. In W-LAN, packets of data are converted into radio waves that are sent to other wireless devices or to a wireless access point (AP)–client connection from the wired LAN to the mobile user. The AP can reside at any node on the wired network and acts as a gateway for wireless users' data routed to the wired network. W-LANs require special MAC layer protocols due to the broadcast nature of radio communication (Chen, 1994). A detailed discussion of W-LAN is beyond the scope of this chapter. W-LANs have gained strong popularity lately and are used widely in health care, industry, commerce, warehousing, and academia. An important feature of the W-LAN is that it can be used independent of a wired network. That is, it can be used as a stand-alone network anywhere to link multiple computers together without extending a wired network. W-LAN uses one of the three basic transmission protocols, namely, direct sequence spread spectrum (DSSS), frequency hopping spread spectrum (FHSS), or low-power narrowband. The majority of RF-based W-LANs operate in the industrial, scientific, and medical (ISM) frequency bands, which are located at 902 to 928 MHz, 2.4 to 2.483 GHz, and 5.725 to 5.85 GHz, respectively. The different architectures of W-LAN based on (Agrawal & Zeng, 2003) are summarized in Table 2.

WLL is a system that connects telephone subscribers to the public switched telephone network using radio waves (International Engineering Consortium, 2003b). With WLL, the traditional copper wire-providing link between the subscriber and the local exchange is replaced by a wireless RF network. WLL is advantageous for remote areas where the cost of wire would be prohibitive, i.e., adverse terrain or widely dispersed subscriber areas. With WLL new service providers can quickly deploy wireless networks to rapidly meet the customer's telephony needs. Existing landline operators can extend their networks using WLL. Cellular telephone companies can deliver residential service using WLL without going through the local telephone company.

WAP is an application environment and set of communication protocols (application, session, transaction, security, and transport layers), which allow wireless devices easy access to the Internet and advanced telephony services (Wireless Application Protocol, 2000; Stallings 2002). WAP offers the ability to deliver an unlimited range of mobile services to subscribers, independent of their network, manufacturer, vendor, or terminal. With WAP, mobile subscribers can access information and services from wireless handheld devices. WAP is based on existing

Internet standards such as the Internet protocol (IP), extensible markup language (XML), hypertext markup language (HTML), and the hypertext transfer protocol (HTTP) and is designed to work with all wireless network technology. More information can be obtained from the *WAP Forum* (Wireless Application Protocol, 2000) and in the chapter on WAP in this encyclopedia.

Bluetooth is a wireless technology that makes possible connectivity to the Internet from mobile computers, mobile phones, and portable handheld devices without the need for cable connections. It facilitates fast and secure transmission of both voice and data, without LOS propagation. Some characteristics of Bluetooth technology are summarized in Table 2. Detailed information on Bluetooth can be found in another chapter in this encyclopedia.

Satellite communication is one of the traditional RF wireless communication systems. Signals can be transmitted directly from a ground station (GS) or gateway on earth to a satellite, and back to another GS. Sometimes, the signal can be routed through another satellite (intersatellite) before it is transmitted back to the GS. We can identify a satellite system by how far the satellite is from the earth. The closer the satellite is to the earth, the shorter the time it takes to send signals to the satellite. There are three satellite orbits, namely, low earth orbit (LEO), medium earth orbit (MEO), and geosynchronous earth orbit (GEO).

LEO satellites are closest to the earth, beginning about 100 miles above the surface, and only take a couple of hours to circle the earth. Because LEO systems are orbiting so quickly, multiple satellites are required to provide constant coverage in one location. LEO systems have the capability to receive calls from the earth and pass them to an earth-based switching system in much shorter time than other satellites. However, because of the speed of the satellite, it is frequently necessary to handoff a particular call to a second satellite just rising over the horizon. This is similar to a cellular mobile radio system (discussed below), except that in this case it is the cell site (the satellite) that is moving rather than the user. The lower orbit has the advantage of allowing access to very low-power devices (Printchard, 1993). LEO satellites are used mainly for wireless transfer of electronic mail, pager systems, worldwide mobile telephony, spying, remote sensing, and video conferencing.

GEO satellites circle the earth at a height of 22,300 miles, orbiting at the same rate as the earth rotates so that they appear stationary from the earth's perspective. Most GEO satellites rely on passive bent-pipe architecture so that they receive signals from transceivers on earth, amplify them, and send them back to specific regions on earth. GEO systems are used for a wide array of services including television broadcasts, long-distance telecommunications, and various scientific and military applications. GEO satellites are well suited to transmitting data, but may be undesirable for voice communications because of the long propagation delay. It takes about one-fourth of a second for a signal to travel from a terrestrial GS to the satellite and back. If the receiver GS replies, it takes another one-fourth of a second, resulting in a total of half a second (Printchard, 1993). This is an unacceptably long delay for voice communication. Hence, voice communications are seldom carried via GEO satellites.

MEO satellites can be found between 1,000 and 22,300 miles and are mainly used for global positioning and navigation systems. MEO satellites are not as popular as the LEO or GEO for reasons beyond the scope of this paper.

New wireless or cellular mobile radio technologies are classified under emerging wireless technologies. These are technologies currently under research and development or technologies that are undergoing field tests. In short, these technologies are not widely deployed. These include the third generation (3G) technologies and the forth generation (4G) technologies. The goal of these technologies is to seamlessly integrate a wide variety of communication services such as high-speed data, video, and multimedia traffic as well as voice signals. Some of these technologies can be realized by combining existing technologies. For example, one of the most promising approaches to 3G is to combine a wideband code division multiple access air interface with the fixed network of a global system for mobile communications (GSM). It is expected that these new technologies will increase the performance of the existing wireless systems. These technologies will provide multimedia capability at much higher rates with Internet connectivity.

RADIO WAVE PROPAGATION

Propagation is the process of wave motion, which is very important in the design and operation of RF systems. Because the received signal is always different from the transmitted signal, due to various propagation impairments, and because of the nature of the propagation itself, it is necessary to understand the properties of radio wave propagation. This is most important in telecommunication applications in predicting the transmission characteristics of the channel. When radio waves are radiated from an antenna, propagation is governed by the following mechanisms.

Free Space Propagation

This is the ideal propagation mechanism when the Tx and the Rx have direct LOS and are separated by a distance d between the Tx and the Rx. If P_t is the transmitted power, the received power P_r, a function of distance d, is given by

$$P_r(d) = P_t G_t G_r \frac{\lambda^2}{(4\pi d)^2 L} = P_t A_{et} A_{er} \frac{1}{(\lambda d)^2 L} \qquad (1)$$

where A_e, G, and L are the effective area of antenna, antenna gain, and system loss factor, respectively. The subscripts "t" and "r" refer to the transmitter and receiver respectively. From this relationship, we observe that the received power diminishes at the rate of 20 dB/decade as the distance increases. The product $P_t G_t$ is defined as EIRP, introduced earlier; i.e., $EIRP = P_t G_t$.

Reflection

When a radio wave strikes an object with dimensions very large compared to its wavelength, reflection occurs.

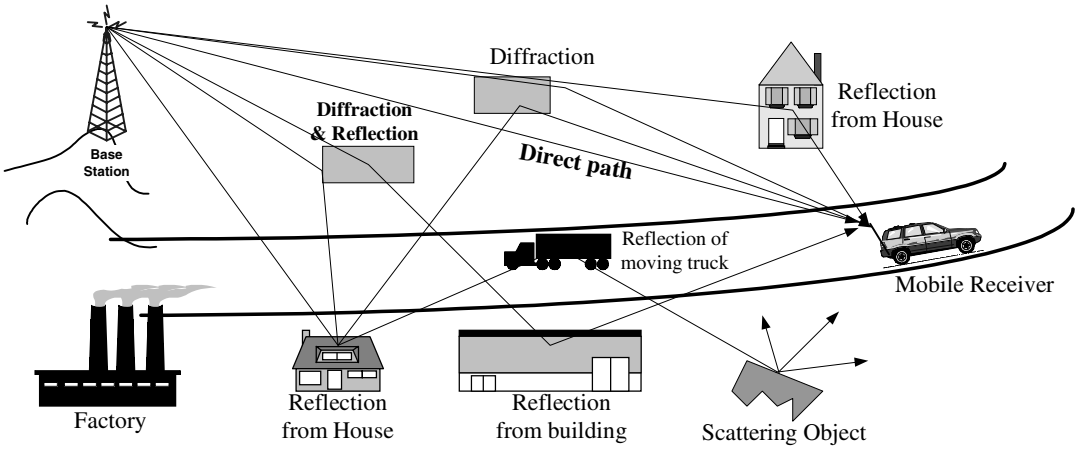


Figure 5: Illustration of reflection, refraction, diffraction, scattering, and absorption.

All radio waves will undergo reflection if the propagation medium undergoes abrupt changes in its physical properties. This is illustrated in Figure 5. The more abrupt the discontinuity, the more pronounced the reflection. Depending on the type of object, the RF energy can be partially reflected, fully reflected, or absorbed. It is possible to compute the amount of reflection from the properties of the two media. This is known as the reflection coefficient, $\Gamma = (\eta_2 - \eta_1)/(\eta_2 + \eta_1)$, where η_1 and η_2 are the intrinsic impedance of the two media. Note that depending on the values of η_1 and η_2, there could be partial reflection, full reflection, or no reflection at all. If the incident object is a good conductor, the wave is totally reflected and the angle of incidence is the same as the angle of reflection.

Refraction

Refraction (see Figure 5) occurs at the boundary between two dielectrics, when the incident wave propagates into another medium at an angle. When radio waves propagate from a medium of one density to a medium of another density, the wave speed changes. This change in speed will cause the wave to bend at the boundary between the two media. The wave will always bend toward the denser medium.

Diffraction

Diffraction of radio waves occurs when the waves encounter some obstruction along their path and tend to propagate around the edges and corners and behind the obstruction. This is illustrated in Figure 5. The height or dimension of the obstruction has to be comparable to the *wavelength* of the transmission. The same obstruction height may produce lower diffraction loss at higher *wavelength* than at lower *wavelength*. The result of this effect is that the object shadows the radio wave. The field strength of the wave decreases as the receiver moves deeper into a shadowed region.

Scattering

Scattering is also illustrated in Figure 5. It is due to small objects and irregularities in the channel, rough incident surfaces, or particles in the atmosphere. When the radio wave encounters objects or particles with dimension smaller than the *wavelength* of the wave, scattering occurs, which causes the signal to spread in all directions.

Interference

Interference can occur when the transmitted radio wave arrives at the same location via two or more paths (multipath). One of the ways this can happen is illustrated in Figure 6. This figure shows three waves arriving at a mobile receiver (the car) after traveling slightly different paths. Due to their phase differences, the radio waves can add either constructively or destructively at the receiver. If the phase shift experienced by the propagating waves

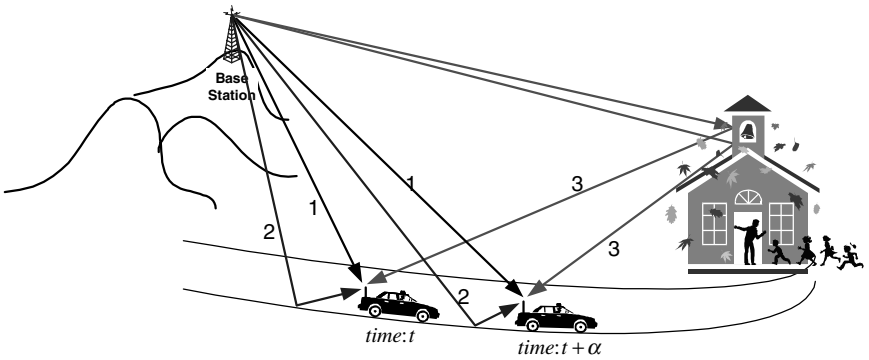


Figure 6: Interference of radio wave.

is time-varying, then it can cause a rapid variation in the received signal, resulting in fading.

Absorption

Absorption describes the process where radio energy penetrates a material or substance and gets converted to heat. Two cases of absorption of radio waves are prevalent. One occurs when radio waves are incident upon a lossy medium and the other is due to atmospheric effects. When the radio wave strikes an object, the incident wave (perpendicular wave) propagates into the lossy medium and the radio energy experiences exponential decay with distance as it travels into the material. The wave either is totally dissipated or will reemerge from the material with a smaller amplitude and continue the propagation. The *skin depth*, δ, is the distance for the field strength to be reduced to 37% of its original value—the energy of the wave is reduced by 0.37. Particles in the atmosphere absorb RF energy. Absorption through the atmosphere also depends on the weather conditions—fair and dry, drizzle, heavy rain, fog, snow, hail, etc. Usually, the absorption of RF energy is ignored below 10 GHz.

Doppler Effect

Doppler shift is the change in frequency due to the difference in path length between two points in space. It is observed whenever there is relative motion between the Tx and the Rx. For a mobile moving with a constant velocity v, the received carrier frequency f_c will be shifted by the amount

$$f_d = f_m \cos\theta = \frac{v\cos\theta}{\lambda} = \frac{v_{eff}}{\lambda} = \frac{v_{eff} f_c}{c} \qquad (2)$$

where θ is the path angle; $f_m = v/\lambda$ is the maximum Doppler frequency f_d, at $\theta = 0°$; and v_{eff} is the effective velocity of the mobile (Garg & Wilkes, 1996). The Doppler shift, bounded by $\pm f_m$, is related to the phase change $\Delta\theta$ caused by the change in path length. Because each component of the received multipath signal arrives from a different direction, each contributes a different value to the Doppler spreading. This effectively increases the bandwidth of the received signal. Depending on the direction of motion and the source, the frequency can be shifted up or down, i.e., $\pm f_m$. The result of this shift is a random phase and frequency modulation of the received RF carrier, which may necessitate the use of differential phase and frequency detection techniques.

The above propagation mechanisms strongly influence system design parameters such as the choice of transmitting and receiving antennas, Tx powers, modulation techniques, and much more. Each of these propagation mechanisms contributes to losses in the RF energy and hence limits system performance. In wireless mobile communications, propagation losses are commonly classified into path loss, shadowing, and multipath fading. These losses are described below.

Path Loss

Path loss (PL) refers to the large-scale envelope fluctuation in the radio propagation environment, which varies with the distance between the Tx and Rx. Because the Rx is located at some distance d from the Tx, a loss factor is used to relate the transmitted power to the received power. For amplitude fading, an increase in d normally results in an increase in PL. Different models have been used to model path loss, but each model obeys the distance propagation law. In free space, with L = 1, PL is expressed as the ratio of the radiated power P_t, to the received power P_r and is given by

$$PL(dB) = 10\log_{10}\frac{P_t}{P_r} = -10\log_{10}\left[\frac{G_t G_r \lambda^2}{(4\pi)^2 d^2}\right] \qquad (3)$$

Shadowing

Due to topographical variations along the transmission path, the signal is diffracted and the average power of the received signal is not constant. Shadowing or large-scale fading refers to slow variations in the local mean of the received signal strength. This variation causes shadowing. The signal is shadowed by obstructions such as buildings and natural terrain, which leads to gradual variations in the mean power of the received signal. The effect is a very slow change in the local mean signal, say P_s. Shadowing is generally modeled by a lognormal distribution, meaning that $s_d = 10\log_{10} P_s$ is normally distributed, with s_d given in dB (Yacoub, 1993). Shadowing is the dominant factor determining signal fading.

Multipath Fading

The collective effect of reflection, refraction, diffraction, and scattering leads to multipath propagation. Due to reflection, refraction, and scattering of radio waves along the channel by manmade structures and natural objects along the path of propagation, the transmitted signal often reaches the receiver by more than one path. This results in the phenomenon known as multipath fading. The signal components arriving from indirect paths and a direct path (if it exists) combine at the receiver to give a distorted version of the transmitted signal. These radio waves are attenuated differently and they arrive with different path gains, time delays, and phases. The resultant signal may vary widely in amplitude and phase depending on the distribution of intensity and relative propagation in time of wave and bandwidth of the transmitted signal. The number of paths may change drastically when the mobile unit changes its position depending on the increase or decrease in the number of intervening obstacles. Unlike shadowing, multipath fading is usually used to describe small-scale fading or rapid fluctuation in the amplitude of a radio signal over a short period of time or over short distances. It is affected by rapid changes in the signal strength over short distances or time intervals and random frequency variations due to varying Doppler shifts on different multipath signals (Rappaport, 2002).

The loss factor associated with multipath fading is usually modeled in the channel impulse response. A transmitted impulse will arrive at the Rx as the sum of several impulses with different magnitudes, delays, and phases. For M multipath, the composite impulse response $h(t,\tau)$

for any given locations of the Tx and Rx is given by

$$h(t, \tau) = \sum_{k=1}^{M} \alpha_k(t)\delta\,(t - \tau_k(t))\, e^{-j\phi_k(t)} \tag{4}$$

where $\alpha_k, (t)$, $\tau_k(t)$, and $\phi_k(t)$ represent the time-varying amplitude, delay, and phase of the kth path signal. This shows that in general, the received signal is a series of time-delayed, phase-shifted, attenuated versions of the transmitted signal. The variables $h(t, \tau)$, $\alpha_k(t)$, $\phi_k(t)$, and $\tau_k(t)$ are also random.

WIRELESS COMMUNICATION TECHNIQUES

Because the wireless channel is not a reliable propagation medium, techniques to achieve reliable and efficient communication are necessary. In mobile channels, for example, the Rx has to constantly track changes in the propagation environment to ensure optimal extraction of the signal of interest. As the receiver moves, the surrounding environment changes, affecting the received signal's amplitude, phase, and delay. The multipath received signals are combined at the antenna either constructively or destructively. During destructive combining the received signal may not be strong enough to produce reliable communication because of the degradation in the signal-to-noise ratio (SNR). It is not uncommon in shadowed signals for the amplitude of the received signal to drop by 30 dB or more within a distance of a fraction of a wavelength (Eng, Kong, & Milstein, 1996). Hence, achieving reliable communication over a wireless channel is a daunting task.

To counter this problem, techniques have been developed for efficient wireless communication. These include spread spectrum, multiple access, diversity, equalization, coding, and related techniques such as multicarrier modulation, orthogonal frequency division multiplexing, multicode and multirate techniques, and multiple input multiple output system, to mention just a few. All these techniques are aimed at increasing the reliability of the channel and the performance of the system. Discussion of some of these techniques is beyond the scope of this paper. However, a summary of the major wireless communication techniques is given below.

Spread Spectrum

Spread spectrum (SS) is a modulation technique where the transmitted bandwidth B_{ss} is much greater than the data bandwidth B_s. The idea is to transform a signal with bandwidth B_s into a noise-like signal of much larger bandwidth B_{ss}. Spreading is usually achieved by modulating the data with a pseudo-random noise (PN) sequence called the "chip" at a rate that is much higher than the data rate. The significance of SS is evident from the capacity equation, given by

$$C = B\log_2(1 + SNR) \tag{5}$$

where C is the channel capacity in bits and B is the bandwidth in hertz. Observe that by increasing the bandwidth B, we may decrease the SNR without decreasing the capacity and, hence, the performance.

The main parameter in SS systems is the processing gain, G_p, defined as

$$G_p = \frac{Spread\ Bandwidth}{Information\ Bandwidth} = \frac{B_{ss}}{B_s} = \frac{T_b}{T_c} \tag{6}$$

where T_b and T_c are the bit period and the chip period, respectively. G_p is sometimes known as the "spreading factor" (Rappaport, 2002). From a system viewpoint, G_p is the performance increase achieved by spreading. It determines the number of users that can be allowed in a system, and hence the amount of multipath reduction effect. It is used to describe the signal fidelity gained at the cost of bandwidth. It is through G_p that increased system performance is achieved without requiring a higher SNR. For SS systems, it is advantageous to have G_p as high as possible, because the greater the G_p, the greater the system's ability to suppress interference. SS techniques are used in cellular mobile telephones, global positioning satellites (GPS), and very-small-aperture satellite terminals. The strength of this system is that when G_p is very large, the system offers great immunity to interference.

There are two major methods of SS modulation, namely direct sequence spread spectrum (DSSS) and frequency hopping spread spectrum (FHSS). In DSSS the frequency of the given signal is spread across a band of frequencies as described above. The spreading algorithm changes in a random fashion that appears to make the spread signal a random noise source. FHSS is the repeated switching of f_c from one band to another during transmission. Radio signals hop from one f_c to another at a specific hopping rate and the sequence appears to be random. In this case, the instantaneous frequency output of the Tx jumps from one value to another based on the pseudo-random input from the code generator. The overall bandwidth required for FHSS is much wider than that required to transmit the same information using only one carrier. However, each f_c and its associated sidebands must stay within a defined bandwidth.

Diversity

Diversity is one of the techniques widely used to increase system performance in wireless communication systems. Diversity combining refers to the system in which two or more closely similar copies of some desired signal are available and experience independent fading. In diversity systems, the received signals from several transmission paths, all carrying the same information with individual statistics, are combined with the hope of improving the SNR of the decision variables used in the detection process. Diversity combining techniques could be based on space (antenna), frequency, angle of arrival, polarization, and time of reception (Eng et al., 1996; Yacoub, 1993). For example, in space diversity the transmitted signal is received via N different antennas with each multipath received through a particular antenna. This can be regarded as communication over N parallel fading channels. Diversity reception is known to improve the reliability of the systems without increasing either the transmitter power

or the channel bandwidth. Regardless of the type of diversity used, the signals must be combined and detected at the receiver. A proper combination of the signal from various branches results in improved performance. The method of combining chosen will affect the receiver performance and complexity. The common combining techniques in wireless communication are maximal ratio combining (MRC), equal gain combining (EGC), and selection diversity (SD). In MRC, the received signals from individual paths are weighted and added in such a way as to emphasize more credible signals and suppress less credible ones (Yacoub, 1993). In EGC, the received signals are equally weighted and then combined without regard to the individual signal strength. In SD, the branch with the best or most desirable signal is selected and the weaker ones are ignored.

Multiple Access

Because the RF spectrum is finite and a limited resource, it is necessary to share the available resources between users. Multiple access techniques are the primary means of sharing the resources in wireless systems. These techniques are multiplexing protocols that allow more than a pair of transceivers to share a common medium, which can be achieved through frequency, time, or code, giving rise to three popular techniques known as frequency division multiple access (FDMA), time division multiple access (TDMA), and code division multiple access (CDMA). In FDMA, the whole spectrum is divided into subbands and the subbands are assigned to individual users on demand. The users use the entire channel for the entire duration of their transmissions. If the transmission path deteriorates, the user is switched to another channel. This access technique is widely used in wireless multiuser systems. Instead of dividing the available frequency as in FDMA, the available time is divided into frames of equal duration in the case of TDMA. Only one user is allowed to either transmit or receive in each time frame. The transmissions from various users are interlaced into cyclic time structure. Instead of using frequencies or time slots, CDMA techniques distinguish between multiple users using digital codes. Each user is assigned a unique PN code sequence, which is uncorrelated with the data. Because the signals are distinguished by codes, many users can share the same bandwidth simultaneously; i.e., signals are transmitted in the same frequency at the same time.

CELLULAR COMMUNICATION

Currently, cellular mobile communication is undoubtedly the most popular RF wireless communication system. In cellular systems, instead of using a single large coverage area with one high-power transceiver (used in traditional mobile systems), the coverage area is divided into small, localized coverage areas called cells. Figure 7 compares the traditional mobile telephone with the cellular telephone structures. Each cell has a base station (BS) or cell site, which in comparison uses much less power. The BS can communicate with mobiles as long as they are within range. To prevent interference, adjacent cells are assigned different portions of the available frequencies. With a certain distance between two cells, the assigned spectrum of a given cell can be reused.

To explain the concept of cellular mobile communication, a summary of the major concepts and techniques is presented below.

Cells

A cell is the basic geographic unit of a cellular system, commonly represented as a hexagon. The term cellular comes from this hexagonal or honeycomb shape of the coverage area. Each cell has a BS transmitting over a cell. Because of constraints imposed by natural terrain and manmade structures, the true shapes of cells are not

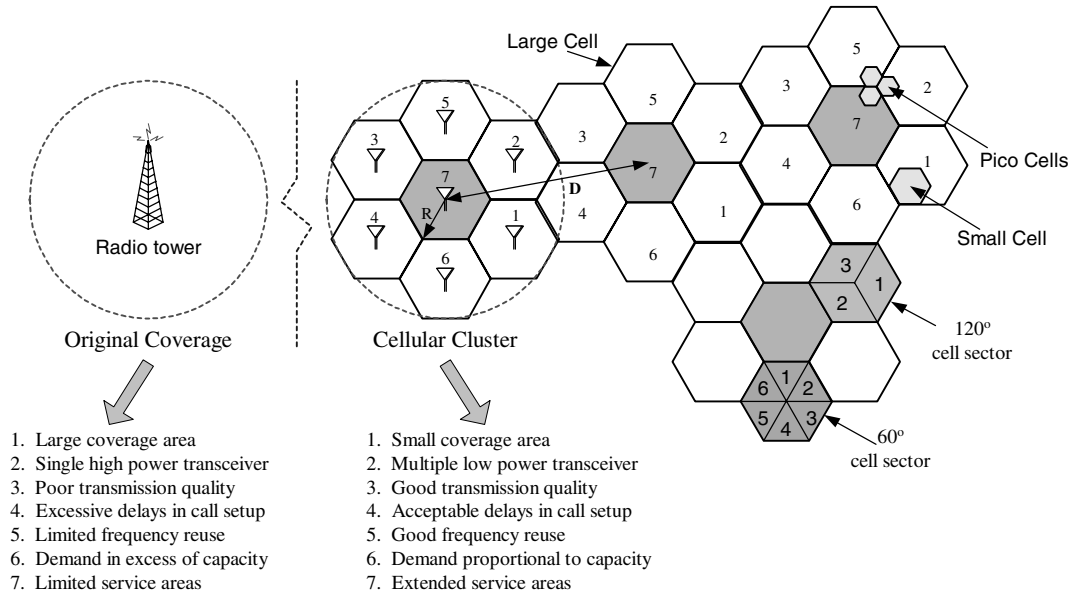


Figure 7: Traditional and cellular mobile radio structure showing frequency reuse, cell splitting and cell sectoring (R = cell radius, D = frequency reuse distance).

hexagons. The coverage area of cells is called the footprint. The BS simultaneously communicates with many mobiles using one channel (pair of frequencies) per mobile. One frequency is for the forward link (BS to the mobile), and the other frequency is for the reverse link (mobile to the BS). Each cell size varies depending on the landscape, subscriber density, and demand within a particular region. Cells can be added to accommodate growth, e.g., creating new cells by overlaying, splitting, or sectoring existing cells. These techniques increase the capacity of the system. Sectoring existing cells and then using directional antennas can also increase capacity.

Clusters

A cluster is a group of cells. No frequencies are reused within a cluster. Figure 7 illustrates a 7-cell cluster, indicated by the dotted circle. Frequency can be reused for all cells numbered 7. Frequencies used in one cell cluster can be reused in another cluster of cells. A larger number of cells per cluster arrangement reduces interference to the system.

Frequency Reuse

Frequency reuse is a technique of allocating channels to cellular systems. Because of the unavailability of spectrum at the cellular band, channel frequencies must be reused. Cells are assigned groups of channels that are completely different from those of neighboring cells. Cells with the same number have the same set of frequencies. If the number of available frequencies is 7, the frequency reuse factor is 1/7, which implies that each cell is using 1/7 of available frequencies (Rappaport, 2002). Frequency reuse introduces interference into the system.

Interference

In cellular mobile communications, there are two types of intrinsic interference, namely, co-channel interference (CCI) and adjacent channel interference (ACI). These interferences are a result of frequency reuse. CCI is the interference between signals having the same frequency (i.e., the reuse frequencies), whereas ACI is the interference between signals having frequencies close together. For example suppose channel 1 has frequencies 825.030 MHz (mobile) and 870.030 MHz (BS) and channel 2 has frequencies 825.060 MHz and 870.060 MHz. Channels 1 and 2 have frequencies close to one another, which will result in ACI. Any other signals having the frequencies of channel 1, 825.030 (mobile) and 870.030 MHz (BS), are co-channel signals and will suffer from co-channel interference. Note that the interference effect is related to the ratio of the reuse distance D and the cell radius R. This is known as the Q-factor Q=D/R and is used to measure the level of CCI. A higher Q value improves transmission quality due to smaller CCI. That is, increasing D improves isolation of RF energy between cells and hence minimizes interference. The ACI is mainly due to imperfect filtering allowing nearby frequencies to leak into the passband of the desired signal (out-of-band interference).

Cell Splitting

Cell splitting is the process of subdividing a congested cell into smaller cells, each with its base station. As the traffic load carried by a large cell reaches capacity, cell splitting is used to increase system capacity. In this way, heavy-traffic regions can be split into as many smaller areas as necessary in order to provide acceptable service levels. Cell splitting decreases R, while leaving Q relatively unchanged. Notice that more cells imply that more cell boundaries will be crossed more often, increasing trunking and handoff. Only those cells that have traffic overloads are candidates for splitting. However, if cells are split in only a part of a system, serious channel assignment problems may result. The difficulty encountered when all the cell sites are not split can be resolved by implementing cell overlay.

Cell Sectoring

Cell sectoring is the process of dividing cells into sectors and replacing a single omni-directional antenna with a directional antenna. Common sectors sizes are 120°, 90°, 60°, and 30°. Cell sectors of 60° and 120° are illustrated in Figure 7. When cells are sectored, R is unchanged, D is reduced, the amount of frequency reuse is increased, and hence capacity is increased. It is observed that the spectral efficiency of the system is enhanced because the frequency can be reused more often.

Handoff

Handoff is the process used to maintain a call in progress when the mobile user moves between cells. Handoff is generally needed in situations where a mobile is at a cell boundary or reaches a gap in signal strength. Because adjacent cells do not use the same frequency, a call must either be dropped or transferred from one radio channel to another when a mobile user crosses the line between adjacent cells. Because dropping the call is unacceptable, the process of handoff is necessary. As the user moves between cells, the transmission is "handed off" between cells in order to maintain seamless service.

EMERGING RF WIRELESS TECHNOLOGIES

The first generation (1G) and second generation (2G) of cellular mobile telephony were intended primarily for voice transmission. This will not be adequate for the new generation of users. With the continued growth of the Internet and World Wide Web, mobile users are continually looking for high-performance wireless Internet technology to enhance their communication capabilities. Although 3G wireless technology has not yet been realized, it promises to enhance users' communication ability, ranging from receiving and sending e-mail to video teleconferencing. The services provided by the generations of wireless technology are summarized in Table 3 (Evolution to 3G/UMTS Service, 2002).

The 3G technologies use wideband code division multiple access (W-CDMA) technology to transfer data over networks. W-CDMA sends data in a digital format over a range of frequencies, which makes the data move faster,

Table 3 Service Differentiation by Wireless Communication Generations

2G	2.5G	3G
Circuit switched	Packet services	Multimedia
Voice	Interactive	High interactivity
Simple message/SMS	Web browsing	Real time road maps
Event notification	E-mail and attachments	Medical imaging
Fax	File transfers	MMS
	Transactions/e-commerce	Audio streaming
	Instant messaging	Video streaming
		Video telephony
	Weighted fair QoS	End-to-end QoS
GSM	GPRS	3GPP (EDGE, UMTS)

but also uses more bandwidth than digital voice services. W-CDMA is not the only 3G technology; competing technologies include CDMAOne, variants of CDMA2000, which differs technically, but should provide similar services.

CONCLUDING REMARKS

RF and wireless communication systems are being used in diverse areas such as home, military, travel, education, stock trading, package delivery, disaster recovery, and medical emergencies. For example, with wireless technology field employees can connect a portable computer via a wireless network to the area office. Sales professionals can stay in touch with customers about products and services, placed orders, status updates to home offices, and inventory. Airline staff can gather information about ticketing, flight scheduling, and luggage using wireless devices. Public welfare agencies such as police, fire safety, and ambulance services can use wireless devices to relay information. Package delivery companies such as Federal Express, UPS, and DHL have adopted the wireless and mobile computing technology for parcel tracking, as well as emergency drop or pickups of shipments.

Although wireless systems are commonplace in our society, the future of the industry is filled with promises and challenges. Future wireless technologies under consideration include 4G mobile technology, multimedia messaging, and complete convergence via broadband, just to mention a few 4G wireless technology based on ultra-wideband communications will enable the use of low-power, high-bandwidth (100–500 Mbps) networks, supporting devices with sense and radar capabilities. Multimedia messaging will allow pictures and sound to be transmitted along with a text message over the mobile phone. Mobile handsets will support full-color display screens, some with embedded Java capabilities, others with digital cameras built in. It is expected that Bluetooth technology will move from theory and hype to practicality, and issues regarding the security of mobile commerce and information security in general will dissipate.

In this chapter, the topics of RF and wireless communication have been discussed. The concept and general definitions are presented. Within these topics, we have discussed the concept of radio waves as propagating electromagnetic waves, including their characteristics and behavior. It is noted that for wireless and mobile radio systems, it is important to understand distinguishing features of the channel, the properties of the radio wave, and several techniques to enhance the reliability of the channel and increase the performance of the system. Also, a summary of the different forms of wireless communication systems was presented, emphasizing cellular mobile radio, which is currently the most prevalent wireless communication system. Finally, the up-and-coming wireless technologies were enumerated. These are the 3G technologies, which will provide more capabilities to their users.

GLOSSARY

1G, 2G, 3G, 4G 1st, 2nd, 3rd and 4th generation wireless systems.

Amplifier An electronic device used to boost the strength of a signal along a communications channel.

Antenna A device used for receiving or transmitting signals.

Bandwidth The capacity of a transmission channel.

Base station Central radio transmitter/receiver that maintains communications with a mobile radio user.

Bluetooth Short-range wireless protocol allowing mobile devices to share information and applications.

Broadband A classification of the information capacity or bandwidth of a communication channel.

CDMA Code division multiple access.

Cellular Wireless communication technique used in mobile phones.

Channel A radio-frequency assignment made according to the frequency band being used.

Downlink Data transmission from a network to a subscriber.

DSSS Direct sequence spread spectrum.

Duplexer Device for isolating transmitter and receiver signals while permitting a shared channel.

EIRP Effective isotropic radiated power: product of power supplied to an antenna and its gain.

FHSS Frequency hopping spread spectrum.

Frequency Rate of signal oscillation in hertz (one hertz is one cycle per second): the number of times a waveform repeats itself in a second.

GPS Global positioning system—a worldwide radionavigation system.

Ground station The ground equipment needed to receive and/or transmit satellite telecommunications signals, including a dish and other electronics components.

GSM Global system for mobile communications, the mobile phone platform used in Europe and many parts of the world.

Handoff Transfer of wireless call in progress from one site to another without disconnection.

Modulation Process of varying a characteristic of a carrier with an information-bearing signal.

PCS Personal communications services: any of several types of wireless voice and/or data communications systems, typically incorporating digital technology.

Propagation Radiation of electromagnetic waves.

Protocol The rules of order by which a communications network is operated.

PSTN Public switched telephone network: a formal name for the landline telephone network.

Receiver Device on a transmission line that converts a signal to whatever type of signal is needed to complete the transmission.

RF Radio frequency: a radio signal.

Spectrum Range of electromagnetic radio frequencies used in signal transmission.

SS Spread spectrum: a communications technology where a signal is transmitted over a broad range of frequencies and then reassembled when received.

Subscriber A cellular telephone user.

TDMA Time division multiple access: a digital communication technology used by some carriers to provide service.

Transmitter The source or generator of any signal on a transmission medium.

Uplink Data transmission in the direction from the subscriber to the network (back to the provider or Internet provider).

WAP Wireless application protocol: a technology designed to provide users of mobile terminals with limited access to the Internet.

Wavelength Distance between points of corresponding phase in two consecutive cycles of a wave.

WLL Wireless local loop: a wireless system meant to bypass a local landline telephone system.

CROSS REFERENCES

See *Bluetooth™—A Wireless Personal-Area Network; Digital Communication; Propagation Characteristics of Wireless Channels; Wireless Application Protocol (WAP); Wireless Communications Applications.*

REFERENCES

Evolution to 3G/UMTS service (2002, August). White Paper. Retrieved January 29, 2003 from http://www.umts-forum.org/servlet/dycon/ztumts/Live/en/units/ResourcesPapers_index/

Acosta, R. (1999). Rain fade compensation alternatives for Ka-band communication satellites. *NASA Technical Memo*. 107534. Cleveland, OH: NASA Glenn Research Center.

Acosta, R., & Horton, N. (1998). V-band and W-band propagation campaign at NASA Lewis. White paper. Cleveland, OH: NASA Glenn Research Center.

Agrawal, D., & Zeng, Q. (2003). *Introduction to wireless and mobile systems*. Pacific Grove, CA: Brooks/Cole.

Bing, B. (2000). *High-speed wireless ATM and LANs*. Boston: Artech House.

Cellular communications services in the USA (2003). Retrieved January 29, 2003, from http://worldofinformation.safeshopper.com/40/778.htm?539

Chen, K. (1994). Medium access control of wireless LANs for mobile computing. *IEEE Network, 8*(5), 50–63.

Durgin, G. (2003). *Space–time wireless channels*. Upper Saddle River, NJ: Prentice–Hall.

Eng, T., Kong, N., & Milstein, L. (1996). Comparison of diversity combining techniques for Rayleigh fading channels. *IEEE Transactions on Communication, 44,* 1117–1129.

Federal Communications Commission (1997). Millimeter wave propagation: Spectrum management implications, FCC Bulletin No. 70. Washington, DC: Federal Communications Commission. Also available at http://www.fcc.gov/oet/info/documents/bulletins/#70

Garg, V., & Wilkes, J, (1996). *Wireless and personal communications systems*. Englewood Cliffs, NJ: Prentice–Hall.

Geier, J. (1999). *Wireless LANs: Implementing interpretable networks*. Indianapolis, IN: Macmillan Technical Publishing.

Goodman, D. (1997). *Wireless personal communications systems*. Reading, MA: Addison Wesley.

International Engineering Consortium (2003a). Personal communications service (PCS). Retrieved January 29, 2003, from http://www.iec.org/online/tutorials/pcs/index.html

International Engineering Consortium (2003b). Wireless local loop (WLL). Retrieved January 10, 2003 from http://www.iec.org/online/tutorials/wll/

Ippolito, L. (1989). *Propagation effects handbook for satellite systems design*. NASA Reference Publication 1082(04). Cleveland, OH: NASA Glenn Research Center.

Mark, J., & Zhuang W. (2003). *Wireless communications and networking*. Upper Saddle River, NJ: Prentice–Hall.

Mobile telephone history (2002). Retrieved December 12, 2002, from http://www.privateline.com/PCS/history4.htm

National Aeronautics and Space Administration (1998). *Systems handbook—Advanced communications technology satellite*. Technical Report TM-101490. Cleveland, OH: NASA Glenn Research Center.

Printchard, W., Suyderhoud, H., & Nelson, R. (1993). *Satellite communication systems engineering* (2nd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Proakis, J. G., & Salehi, M. (2002). *Communications systems engineering* (2nd ed.). Upper Saddle River, NJ: Prentice–Hall.

Rappaport, T. (2002). *Wireless communications: Principles and practice* (2nd ed.). Upper Saddle River, NJ: Prentice–Hall.

Selected U.S. radio frequency allocations and applications: 70 MHz to 700 MHz, 1997 (2002). Retrieved November 2002, from http://www.rfm.com/corp/new868dat/fccchart.pdf

Stallings, W. (2002). *Wireless communications and networks*. Upper Saddle River, NJ: Prentice-Hall.

Weisman, C. (2003). *The essential guide to RF and wireless* (2nd ed.). Upper Saddle River, NJ: Prentice-Hall.

Wenig, R. (1996). *Wireless LANs*. Boston: Academic Press.

Wireless application protocol (2000, June). White paper. Retrieved December 6, 2002 from http://www.wapforum.org/what/WAP_white_pages.pdf

Wireless communications, market & opportunities (2003). Retrieved January 29, 2003 from http://www.igigroup.com/st/pages/chinav4.html

Xu, H., Rappaport, T., Boyle, B., & Schaffner, J. (2000). Measurements and models for 38-GHz point-to-multipoint radiowave propagation. *IEEE Journal on Selected Areas in Communications, 18*(3), 310–321.

Yacoub, M. (1993). *Foundations of mobile radio engineering*. Boca Raton, FL: CRC Press.

Real Estate

Ashok Deo Bardhan, *University of California, Berkeley*
Dwight Jaffee, *University of California, Berkeley*

INTRODUCTION

The diffusion of major new technologies impacts the economy in many ways. Their widespread adoption affects the way things are produced, distributed, and consumed. New technologies influence business organization, market structure, and productivity, among other economic variables. The Internet is no exception. Its convenience, speed, low cost, and versatility are being exploited on a daily basis in ever-changing ways. Industries that have been functioning for decades in a traditional manner, such as real estate, have also become major targets of the transforming power of the Internet.

E-commerce has the capability of transforming the sectoral structure of the economy by creating new industrial sectors or subsectors, by replacing existing, traditional sectors and by changing the mix and range of services provided. Real estate (broadly defined to include builders, brokers, real estate services, and real estate finance and investment) illustrates all the mechanisms through which a traditional industry undergoes rapid transformation under the impact of the Internet. The Internet, including its economic progeny e-commerce, has already influenced the functioning of the real estate industry in direct and indirect ways. It has become a marketing and sales tool that allows a real estate business a greater reach than before. It has affected the location decision—where and how firms do business—which in turn determines the role of firms involved in brokerage, real estate development, investment, and finance. Reductions in transaction costs, coupled with the qualitatively different nature of information dissemination and communications are the primary channels through which the Internet impacts the basic structure and operation of real estate markets. The impact of e-commerce on the provision of real estate services can therefore occur through changes in either the cost or the type of services. That is, e-commerce can do the following:

Reduce the cost of providing traditional real estate services;
Expand the range, form, and content of traditional services; and
Create new value-added services.

THE INTERACTION OF E-COMMERCE AND REAL ESTATE MARKETS

The interaction of e-commerce and real estate markets involves three sets of features. First, there are technological characteristics of the Internet that have an impact on real estate related activity. Second, there are general features of the Internet and e-commerce that make its application to real estate markets especially valuable. Third, there are specific features of real estate markets that benefit in special ways from the use of e-commerce. This section considers these three forms of interaction.

Internet Attributes of Particular Relevance to Real Estate

Together with other sectors of the economy, real estate shares in some of the basic advantages of the Web, such as ease of marketing, communication, and feedback from clients, lowered costs of operations and sales, and convenience of customer service and support. Indeed, because of the dispersed, localized nature of information in real estate, the prospective gains from information dissemination, ease of comparability, and Web links are particularly significant for real estate firms (see Table 1).

Table 1 Internet Characteristics and Real Estate Firms

Internet/Web Feature	Type of Real Estate Firm
Search Capability; Graphics, Virtual Tours	*Listing/Matching & Brokerage Online Mortgage*
Online Communities & Markets, Multilateral Coordination	*Project Development, Project Finance*
Interactive Transactions Capability	*Property Management*
Online Tracking, Database Access and Analysis	*Consulting Firms; REITS*

In addition, the Web possesses three features that are of particular relevance to the real estate industry:

1. **Graphics/visualization capability**: In its most state-of-the-art form, Web sites now allow prospective buyers and renters to take virtual tours of homes, resorts, hotels, and convention centers. Virtual tours are designed with the objective of bringing about a decrease in the number of properties physically visited before the final sale or rental.
2. **Increased geographical reach**: A unique feature of the real estate industry is that real estate is, of course, locationally specific. A physical and local presence, therefore, has generally and traditionally been critical for firms providing services for real estate transactions. E-commerce has the potential to reduce and in some cases even eliminate the need for a local presence. For example, property sales data were previously available primarily in hard copy prior to e-commerce, requiring a local physical presence for access. Large amounts of these data are now in electronic databases, making access independent of location. Not only service providers, but also buyers and sellers of a specific property, can now be represented electronically, without a physical presence. This affects the structural composition of the industry as a whole and impacts individual firms and their organizational makeup. Although the "local" aspect of real estate will perhaps never be whittled away completely, there is no doubt that inquiries about properties can now emanate from far away to a much greater degree than before. This, in turn, potentially increases the market's size and depth, making it more efficient.
3. **Collaborative and interactive features**: The establishment of new kinds of firms in the real estate sector, which deal with development and predevelopment sites, has been made possible by the multilateral, collaborative platform provided by some Web sites. Large, complex projects create logistical and coordination needs of a wide range of professionals, such as architects, engineers and subcontractors who collaborate and interact on a continuous basis. These needs are met by firms that provide Web sites acting as information clearinghouses, and through which all the participants in the project can coordinate their activities and keep abreast of the latest changes in plans, budgets, and blueprints. The Web-attributable features that make this kind of value addition possible are instantaneous comparability, interactive capability, online calculations and communication, and efficient information management.

General Features of E-commerce and Their Relevance for Real Estate Markets

Reductions in transaction costs are the most important general feature of E-commerce for real estate markets. As a closely related matter, new services will emerge where it was previously not cost effective to provide them at all. The two most important economic aspects of e-commerce that lead to reductions in transaction costs are network externalities and economies of scale; see Shapiro and Varian (1999) for a general discussion of Internet economics.

Network Externalities

Network externalities occur when the value of a network to each user rises as the total number of users increases. A telephone network is an example of this demand-side economy of scale. A telephone has limited value if only a small number of people are connected to the phone network. But as the number of people with connections rises, the value of the telephone network to each one increases. While wires physically connect the nodes on a traditional telephone network, there are *virtual networks* as well. Virtual networks, and the related membership group of a *virtual community*, are key mechanisms through which e-commerce creates value for real estate markets.

A good example of the need for a real estate virtual community is the process of selling residential properties in a particular community. Efficiently matching buyers and sellers in such a market is a complicated, but obviously critical, activity. Traditionally, it has required the physical presence of buyers, sellers, and a real estate agent who brings them together. The physical presence of these parties raises the costs in terms of both time and money. E-commerce, however, can readily create a virtual community, in which each potential buyer sorts through the features of the available homes, and then takes a virtual tour of the most promising homes. In this way, e-commerce can eliminate a significant proportion of the search costs.

Another example of a real estate virtual community consists of all the participants in a large-scale commercial

real estate construction project. These participants include developers, architects, city permit inspectors, lenders, investors, workers, material suppliers, to name just a few. For the construction process to proceed smoothly, it is critical that all these participants be aware of what all the other participants are doing on a real-time basis. It is equally clear that high-cost mistakes can readily occur if such communication fails to occur. E-commerce allows the creation of a virtual community in which all these participants are members, and through which instantaneous communication, coordination, and collaboration can occur at very low cost.

Economies of Scale

Database services often exhibit economies of scale since, as the volume of users rises, the fixed cost of creating the database can be spread over a larger number of users. One example of such a database in real estate markets is the set of information that can be distributed to potential homebuyers in a specific community concerning the characteristics of that community, such as schools, parks, and other public services. The delivery of such information can occur through a virtual community, showing that in the application of e-commerce to real estate both network externalities and economies of scale often interact, increasing the benefit of each one.

Specific Interactions of E-commerce with Real Estate Markets

E-commerce is especially valuable for real estate because real estate markets use a great deal of *intermediation resources,* the technical term for the brokers and agents who expedite real estate transactions. Examples include real estate brokers (bringing buyers and sellers together), mortgage brokers (bringing borrowers and lenders together), and insurance agents (bringing insurance companies and their customers together). These intermediaries provide valuable services, which is why real estate market participants are willing to pay their fees. At the same time, the intermediaries are only a means to an end—to complete a real estate transaction—and the market would function better if their services could be provided in a more efficient manner. Underlying the potential of e-commerce to provide such intermediation services in a more efficient manner are the two factors discussed in the previous section, network externalities and economies of scale.

Real estate markets tend to maintain a stock of vacant units, since the matching of demand and supply is always imperfect and takes time. Therefore, the effect of a more efficient process of intermediation between demand and supply is likely to be a reduction in the long-run, natural, vacancy rate. This reduction in the number of vacant units is a social benefit, since it means these resources can now be put to some other use.

REAL ESTATE MEETS THE INTERNET

Real estate firms and related businesses were among the early private sector pioneers of Internet use and have had a fast growing presence on the Web. One example of the real estate sector's presence on the Internet in its pre-World Wide Web incarnation was the real estate classified bulletin board of Prodigy, the online service which had listings for homes and other real estate. Some real estate related Web sites started in 1994. For example, the New York City Real Estate Guide Web site, created in the summer of 1994, was one of the first to offer free access to the latest New York real estate information. By the summer of 1995, the site was receiving more than 100,000 inquiries a month.

Table 2 Stages of Real Estate Web Sites

- **The Informational or "Presence Signaling" Stage**
  Staking Claim to space on the Internet; Basic Information Dissemination
- **The Brochure or "Marketing/Advertising" Stage**
  Detailed Information; Customer Service; Internet as Marketing Tool
- **E-commerce or "Cyberspace Office/Store" Stage**
  Transactions on Web; Full Service Office in Cyberspace
- **Multilateral Collaborative Stage (For Some)**
  Multilateral Coordination/Collaboration in a Virtual Community and Interfirm Linkages for Complex Production Activity

By the end of 1995 there were close to 4,000 real estate Web sites. The content of these sites, as well as the mix of real estate related firms on the Web, has changed over time. Initially, quite a few of the sites were residential real estate brokerages and listing guides, but fairly rapidly the list expanded to include commercial and retail listings, mortgage brokers, appraisers, architects, real estate attorneys, developers, construction firms, and suppliers. As investment vehicles for real estate expanded, real estate investment trusts (REITs), publicly held firms, and investment advisors also added Web sites. A particularly high proportion of real estate brokers are taking advantage of Web technology. A survey conducted by Real Estate Broker's Insider in their February 1998 issue confirmed that nearly 95% of the respondents/brokers had a Web site, and that even back then, more than 90% of the housing stock on sale at a given time was listed on the Web.

Existing private sector real estate Web sites can be categorized into three, or possibly four, types, as summarized in Table 2; see also Table 3 for different e-commerce models in real estate, such as business-to-business (B2B) and business-to-consumer (B2C).

The most basic sites provide simple information dissemination. The firm registers a Web site and develops a page giving basic company information, signaling its presence on the Internet. The second stage involves using the Internet as a marketing and customer service tool. From a marketing viewpoint, information dissemination and customer services on the Web can be monitored and analyzed differently from conventional methods. Internet

Table 3 E-commerce Models in Real Estate

B2B Web Sites	B2C Web Sites
Project Development	*Online Mortgage*
Supply Management	*Listing/Matching & Brokerage*
Virtual Community/Markets	

tools can now provide a firm with data on who accessed the site, which pages were visited most frequently, from where and for how long. This information contributes to improved measures of the results of promotional efforts. The promotional costs associated with the Internet have also been very low. For example, in direct mail marketing, it is generally more expensive to send a one-page color brochure to 5,000 random addresses than it is to add a component to an existing Web site.

The third stage is represented by a full-fledged office/store on the Web with transactions capability. Some firms have even developed a fourth type of Web site that exploits the multilateral coordination and collaborative capability of the Internet. As mentioned earlier, this type of Web site provides a platform that helps in fine-tuning project requirements, forecasting cost overruns, as well as dealing with logistics. This is a qualitatively new kind of economic activity in real estate, and hence a new kind of Web site, whereas the first three stages of Web sites can be seen as evolving sequentially.

E-COMMERCE AND TYPES OF REAL ESTATE FIRMS

Real Estate Brokerage Services: Disintermediation or Reintermediation?

Real estate transactions have relied heavily on intermediaries. Most obviously, brokers use their specialized knowledge to aid buyers and sellers. E-commerce introduces new opportunities, improving the availability of information, reducing transactions costs, and facilitating the searching and matching process. At the same time, e-commerce has reduced barriers to entry and these opportunities may now be available to new entrants to the industry, thus increasing competition.

Consider a specific community in which there are 100 homes for sale and 100 households potentially interested in buying such a home, the potential buyers being scattered across the United States. A key goal of such a real estate market is to match each home with that household for whom the home is the best possible fit. A physical visit of each household to each house, of course, would be prohibitively expensive, requiring 10,000 ($= 100 \times 100$) house visits. Instead, both the buyers and the sellers hire real estate agents, who have special information regarding the particular market, in order to expedite the process of matching buyers and sellers. Of course, the agents must be paid, and their fees can easily represent 6% of the value of the total transaction. Not only is such intermediation costly, but also it may be imperfect, in the sense that the best matches may go unrecognized. E-commerce can provide a substitute for the traditional real estate agent in at least three forms.

First, Web directories and specific Web pages can be created with properties "for sale by owner," allowing such properties to be listed without the resources of a real estate broker. Properties "for sale by owner," of course, existed through newspaper ads even before e-commerce, but e-commerce provides a much more efficient method for allowing potential buyers to "view" the property and ultimately to complete an actual transaction with the seller. So far, however, "for sale by owner" through e-commerce continues to be a relatively small part of most real estate markets.

Second, the information traditionally maintained by real estate brokers can be distributed much more efficiently in an electronic form across all relevant buyers and sellers. The most important example here is an *electronic multiple listing service*. Traditionally, most real estate markets maintained and updated a hard copy describing all properties for sale in that market. This was described as a multiple listing service, since it represented a cooperative effort of the real estate agents in this market. With e-commerce, the multiple listing service becomes a Web page, with a variety of major advantages, including timely updating, versatile database viewing, selection by parameters and much better graphic displays including virtual house tours. The impact is felt in both shorter, more efficient, search processes and in a smaller number of physical visits.

Third, e-mail has emerged as the alternative means of communication between brokers and their clients, substituting for time-consuming and costly face-to-face meetings, faxes, and snail mail. Constant updates and clarifications are now much more conveniently made at times that are individually and separately convenient to both parties, rather than at moments that are simultaneously convenient to them.

It appears, however, that most buyers and sellers of single-family homes continue to need the services of real-life real estate agents and brokers. These services include advice on listing prices for sellers and offer prices for buyers, individual advice on house attributes (location and quality), and referrals to other experts. Interviews with residential brokers suggest that many have chosen to develop a range of additional services, such as Web links to reliable contractors, to appliance vendors, and to local government agencies. The implication is that the Internet and real-life brokers are likely to be complements, not substitutes, over a broad range of home-buying services.

For commercial and industrial brokers in general, the immediate advantages of the Web are few, while the challenges they face appear greater. The residential brokerage system, as mentioned above, already had a database in place with shared listings, making the transition to a Web-based system of sharing information fairly straightforward. The commercial and industrial sectors, in contrast, had not created any basic systems or databases for sharing information. Despite this initial condition, the commercial/industrial brokerage sector is now a major user of the Web and is helping in the creation of a truly national real estate market in properties for sale, lease, or

rent. Individual brokerage companies have complex Web sites that provide information on the local area as well as nationwide markets. New listing services have developed that allow for database searches, both locally and nationally, for suitable buildings or sites. In addition to nationwide databases of properties, a number of firms offer capability for online rental inquiries, as well as online lease agreements.

The potential of electronic multiple listing services raises, of course, the question of competition between traditional real estate agents and the new instruments of e-commerce. This question is considered below.

Real Estate Auctions

Auctions have long been the center of a dilemma in real estate markets. On the one hand, auctions would seem to be highly efficient mechanisms for selling properties, by giving all potential buyers easy and equal access to the bidding process. For example, auctions are commonly used in Australia and New Zealand (see Dotzour, Moorhead, & Winkler, 1998). In the United States, on the other hand, auctions are rarely used and are clearly dominated by the traditional method of "listing" a property with a broker, and then waiting until the broker finds a buyer willing to pay a sufficiently high price to close the deal. What is going on here?

The answer seems to be that, at least in the United States, the traditional physical auction, taking place at a specific location and a specific time, is too constraining in terms of limiting the number of buyers that can actually participate in the bidding process. Thus, with certain exceptions, sellers find that they obtain a higher expected price using the traditional "listing" method. E-commerce, however, provides the potential to resurrect the auction and make it into a key component of many real estate markets. Electronic auctions, relative to physical auctions, have the key advantages that the auction can be left open for a considerable period of time, certainly a week, and possibly a month, and during this time there is no need for participants to be physically present. Other advances in real estate markets that are provided by e-commerce, such as large, automated databases, further enhance the potential for electronic auctions. Nevertheless, as of this writing, electronic real estate auctions have still failed to capture any significant share of the market for real estate transactions.

E-appraisals

Appraisals play an important role in real estate markets, especially as a means for lenders to determine how much money they are willing to lend to the new buyer. In the past, appraisals have often been costly, time-consuming, and inaccurate, significantly raising the costs of carrying out real estate transactions (see Diaz, 1997). E-commerce in the form of automated appraisals has the potential, however, of significantly improving all aspects of the appraisal process. The key factor here is that information on comparable sales is the raw material for appraisals, and this information can be most readily accessed and applied using modern electronic database techniques. Already, there is extensive use of electronic appraisal techniques by government sponsored mortgage agencies such as Fannie Mae and Freddie Mac, and it seems highly likely that these techniques will come to dominate the entire market. This raises the important question, however, whether the new automated services will be provided by new e-commerce firms, or whether traditional appraisal firms will be the providers of these services.

Real Estate Finance

The premium on rapid dissemination of quality information has made finance a very fertile field for Internet usage. The range of real estate finance related Web sites extends from online mortgage firms to those involved in private project financing and equity placements.

Residential Mortgage Lending

Key features of residential mortgage lending that make it suitable for e-commerce enhancements include

a. The need to reference large databases to perform credit checks on the individual and the appraisal on the property;
b. The timely nature of the credit approval process, since the buyer will want authority to proceed with the purchase as soon as possible; and
c. The large menu of choices for mortgage contracts, and the need to update these daily to reflect the most current market conditions.

From the consumer point of view, online mortgage firms provide an efficient entry point in the search for information on mortgages, rates, and fees; see the survey in Mortgage Bankers Association of America (2001). Mortgage lending, however, is highly regulated at the state level in the United States, with the result that a one-size-fits-all lending platform may well run afoul of various state laws. This means that the economies of scale that might be otherwise be available by scaling a single platform to all 50 states will at least be somewhat limited. The issue here is not whether electronic methods will be used in mortgage lending—virtually all lenders now rely on these methods—but rather whether they will simply be efficient tools used by traditional lenders, or whether they will allow the entry of a completely new set of lenders, the so-called "e-lenders."

Commercial Real Estate Finance

The needs of commercial real estate finance are being served by a host of Web sites that are bringing developers, brokers, investors, and lenders together. Developers and sellers of projects and properties make their presentations and solicit offers on these Web sites. The sites involve variations on interactive meetings facilitating free flow of information, and in some cases incorporate due diligence filtering procedures. Mortgage backed securities sites offer data and information on ratings, duration, spreads, delinquency rates, and upcoming offerings. Borrowing from other, more general, finance Web sites, there are now a plethora of sites catering to potential individual investors in REITs and limited partnerships and on speculative property purchases. E-commerce has the potential

to greatly increase efficiency and depth by expanding access to finance from personal networks to a broader set of virtual, anonymous financial markets.

Property Management

Property management companies with large portfolios of apartments need to integrate specialized property management software on the one hand with effective communication with investors, customers, and managers on the other. Both the Internet and its segmented, private version—the intranet—are now being used for to access data and analysis from these applications and then to disseminate the information. At any given time, it is possible to retrieve the latest data regarding revenues, expiring leases, and vacancies. Real-time online tracking and database accessing are made possible by this two-way transfer of processed information.

Project Management/Development and Predevelopment

The complex coordination needs of large projects are now being met by a new generation of Web sites that combine virtual community creation, online collaboration, and support services to create an environment in which the entire process from the design stage to the construction process is streamlined. The scattered activities of subcontractors, architects, engineers, developers and others can now be brought together on a technology platform under the aegis of an Internet firm in order to facilitate quick changes in blueprints, fine tuning of work in progress, and resolving supply bottlenecks. Some developers, construction firms, and contractors have their own specific, project-linked, intranets to manage their supply-chain issues. Some predevelopment oriented Web sites have developed online technology platforms for studying cost metrics, design parameters, and feasibility analysis. Digital storage of project data and information, accessible databases/blueprints, online updating/fine tuning, supply management, and project planning capabilities result in new kinds of value creation in real estate.

Relocation Services

The United States has an extraordinarily high internal mobility, with more than 20 million people moving each year. A number of Internet-based relocation services firms have appeared offering complete relocation packages (packing, shipping, trucking, etc.), with links to moving companies, listings for the destination point, short-term rentals, and other relevant information about the destination city. Some sites also come with a range of excellent tips and checklists, with both "before move" and "after move" versions. For corporate customers there are modules for cost comparisons, as well as absorption data, vacancy rates, and office, retail, or industrial spaces available.

Title and Other Property Insurance

Insurance plays an important role in all real estate transactions, since both investors and lenders will desire protection against physical risks (such as fire) and the risks of an invalid title (title insurance). Many activities of an insurance company—actuarial, claims, and billing—are highly data and informative intensive. So it is not surprising that insurance firms are taking significant advantage of Internet capabilities in managing their back offices. Furthermore, the Internet is increasingly being used to market insurance, based on the ability of firms to offer policies based on the information that the consumer provides, and for consumers to carry out comparison shopping. Title insurance is a particularly interesting example, since records pertaining to property ownership are now becoming available online. The computerization of these records together with the advent of the Internet is allowing title insurers to expand to serve national markets.

INTERNET USE: THE EXPERIENCE OF REAL ESTATE FIRMS AND CONSUMERS

The Fisher Center for Real Estate and Urban Economics at the University of California, Berkeley, carried out a limited survey of a sample of 60 leading real estate related firms in the United States and California. [The sample consisted of those firms that were members of the Advisory Board of the Fisher Center for Real Estate and Urban Economics at the University of California, Berkeley. These firms cover a wide range of real estate activities and are among the largest in the industry. See Bardhan, Jaffee, & Kroll (2000).] The survey showed that over four-fifths had Web sites by March 1999, with about one-third having already established their sites by the end of 1996 and another third with sites inaugurated in 1998 or early 1999. Brokers, lenders, financial services firms, law firms, and residential developers were among the early adopters. Commercial developers, consultants and advisors, lenders, REITs, and investment firms were among the later adopters. Those without sites were more likely to be privately held firms with a relatively narrow base of activity.

Most firms with Web sites used their site to provide information about the company and to market services. Some marketed property from their site (either as individual pieces or as part of a REIT), providing detailed information on the characteristics of buildings available, surrounding communities, and other related data. Other Web site uses included employee recruiting, providing information for members or investors, and disseminating related information on topics such as regulations or real estate markets (see Figure 1). It should be noted, however, that both the survey mentioned above and one from Georgia Institute of Technology mentioned later will become outdated in a field that is changing as rapidly as e-commerce.

From the point of view of real estate firms, a key feature of the Internet is to create initial leads that are later followed by transactions. Real estate businesses use the Internet initially for marketing and communication, and later additionally for customer support and service.

Web sites frequently lead to contacts that are then nurtured through telephone and person-to-person meetings. For residential real estate, Web activity from the point of view of the consumer includes residential searches,

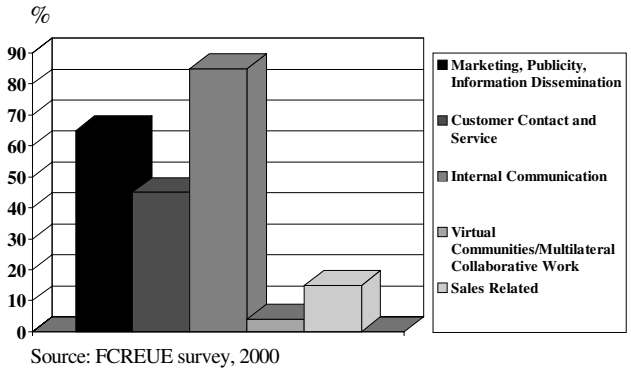


Figure 1: Web use by real estate businesses.

other housing information, and pricing comparisons (both on houses and mortgages), with follow-up contact with brokers. A significant volume of real estate related transactions is seen in the hospitality industry (making reservations for hotels and vacation homes), and to some extent in online mortgage applications (see Table 4 for how consumers and businesses use the Web in real estate transactions).

Surveys carried out by the Graphics Visualization and Usability Center (GVU) at the College of Computing, Georgia Institute of Technology, help us to understand the demographic and other determinants of people who access real estate sites. GVU has been carrying out these online surveys on growth and trends in Internet usage since 1994, and they cover issues of demographics, usage patterns, attitudes, social standing, commercial usage and occupation, among others. The authors have carried out an analysis based on GVUs downloadable raw data of the tenth survey carried out in 1998; see http://www.gvu.gatech.edu/user_surveys/.

Of the range of variables covered in the survey, the variable "Real Estate Access" is of particular relevance. These data were used to estimate a probit equation. Responses involving accessing of real estate sites at any frequency were coded as 1 and not having ever accessed any site as 0. This became the dependent variable. The explanatory variables are education (college and above = 1), gender (male = 1), age (below 30 = 1), and usage pattern variables that were coded in a similar way.

The results of the probit estimation are shown in Table 5 as marginal effects, or transformed probit coefficients. The probability of accessing real estate Web sites increases with college education (4%), youth (< 35 years; 7.5%), household income (> $75,000; 3.8%), and decreases if the Web surfer is male (6%). The probability also increases if Web usage is with the purpose of looking for product information (6%) and for shopping on the Internet (10%). Since the sample has some selection issues, the results can be considered only indicative. Perhaps the coefficient on the age variable is the somewhat surprising, but since the survey did not differentiate between accessing real estate Web sites for purposes of renting, researching, or with a view to purchasing, it is possible that young renters were a sizeable part of the sample. Of the total sample of 3,206 respondents, more than a third (1,226) had accessed real estate Web sites.

Table 5 Marginal Effects on Accessing of Real Estate Web Sites

Independent Variables	Marginal Effect	Standard Errors
College Education	0.04	0.017
Age	0.075	0.017
Male	−0.06	0.018
Searching for Product Information	0.06	0.025
Internet Shopping	0.10	0.017
Income	0.038	0.02

Note: All coefficients, except for the one on the income variable, are significant at the 5% confidence level. Pseudo R2 = 0.08; $N = 3206$; log likelihood = −2087.

CONCLUDING REMARKS: THE EMERGING STRUCTURE OF NEW E-COMMERCE REAL ESTATE MARKETS

Impact on Real Estate Firms

The interaction of traditional real estate markets with e-commerce is having an impact on both real estate firms and on real estate markets. For existing firms, a major adjustment is to incorporate the Internet directly in their operations. At the same time, a new generation of firms is entering the market, some providing traditional services with e-commerce techniques and others providing new services of value in the marketplace.

Table 4 A Real Estate-Oriented View of the Web

Point of View of Real Estate Business	Point of View of Real Estate Consumer
• Marketing/Publicity • Customer Service • Direct, Instantaneous Feedback & Communication • Customer Support • Online Sales	• Product Information Research • Price and Attributes Comparison Communication • Ease & Convenience of Ordering, Payment, Delivery and After-Sales Service

A key question is whether new services will be provided primarily by new, specialized, e-commerce firms, or by already existing firms in the industry. Generally speaking, new e-commerce firms will have better information regarding the new electronic techniques, while the existing firms will have a comparative advantage in information regarding the real estate market itself. The "winners" will be those firms, new or existing, that can combine both areas of competence to create value-added service products for the real estate industry. In this respect, the adjustment of the real estate industry to the Internet is similar to that of established firms in many other economic sectors.

In the areas of real estate brokerage and real estate lending, it appears that traditional firms are likely to continue to dominate the real estate service markets. In the real estate brokerage area, control of new real estate listings and multiple listing services by traditional firms should allow these firms to continue to dominate the market. In the real estate lending area, automated techniques were rapidly adopted by most existing lenders, thus taking the initiative from new e-commerce entrants. In both cases, however, the traditional firms will have to continue to innovate and adopt new electronic techniques to sustain their dominance.

In the areas of real estate appraisals and construction project management, in contrast, new e-commerce firms may well dominate. In the case of real estate appraisals, the existing industry may be too slow to adopt the new electronic techniques, and thus ultimately may lose their market share. An important factor here is that the existing industry consists largely of very small firms, often just individuals for whom it is not feasible to adopt the new techniques. In the case of construction project management, basically a whole new service area is being developed, so by its nature it requires new firms.

There is also the question whether the adoption of e-commerce techniques in real estate service markets will create incentives for mergers across service lines, creating multiline service providers. For example, is the merger of real estate brokers, lenders, and insurers into single mega-sized firms likely? E-commerce can provide incentives for such mergers as a result of the network externalities and economies of scale discussed earlier. Specifically, e-commerce techniques may provide an important means to bundle and cross sell real estate services. On the other hand, regulatory prohibitions are the main reason that existing real estate service providers have not already merged in order to bundle and cross sell their services. Thus, the creation of highly integrated e-commerce real estate service firms would require a significant change in the regulatory environment, something unlikely to happen rapidly.

Impact on Real Estate Markets

Decreasing transaction costs and relatively low barriers to entry have increased competition in some sectors of the industry. Costs have been lowered by shortening the transaction cycle and by precise market targeting, in addition to the savings in information dissemination. A major open issue is whether the Internet and e-commerce will have a significant overall impact on the demand for various types of real estate (see Table 6). There is some mixed evidence on the impact of e-commerce on retail real estate markets. Schlauch and Laposa (2001) reported that retailers who incorporate online operations into their overall real estate strategy manage to lower somewhat their real estate related costs. At the same time, new electronic retailers—that is, retailers who conduct business purely over the Internet—are increasing the demand for warehouse space.

Table 6 Potential Effect of Internet and E-commerce on Real Estate By Type

Category of Real Estate	Through Type of Internet Related Activity
Office	Telecommuting
Residential	Telecommuting
Warehouse	E-commerce
Retail	E-commerce
Hotels/Resorts	Online Reservations

In the long run, the fast growth of electronic retailers and the slower growth of traditional retailers may create a shift in real estate demand from traditional retail space toward warehouses. There might also be a net contraction in total real estate demand, since the electronic retailers will likely require less space per dollar of sales. In the short run, however, the total demand for real estate may actually expand, since traditional retailers may contract very slowly, while the new electronic retailers are expanding very rapidly.

Whatever the details, the fundamental point is that the real estate markets will be most influenced by those firms that can create value at the intersection of real estate and Internet technology.

GLOSSARY

Agent, real estate Another term for real estate broker.

Appraisal, real estate Estimation of market value of a property or home as concluded by a third party, usually a licensed professional.

Broker, real estate An entity or person who brings together potential buyers and sellers of real estate and generally charges a fee for the services. In most states, real estate brokers must pass exams to be licensed. Brokers may also act as intermediaries in leasing transactions.

Commercial real estate Real estate properties used for commercial purposes, including office buildings, shopping malls, and hospitals, among others.

Development, real estate The process of transforming vacant or underused land into residential or commercial real estate.

Economies of scale Situation in which the average production costs fall the greater the volume produced.

For sale by owner Properties that are listed for sale by the owner, without the use of a real estate broker.

Intermediary, real estate A firm or individual who provides valuable services to enable or facilitate a real estate transaction.

Mortgage A financial instrument that is used to borrow funds to enable the purchase of real estate. The property is used as collateral, whereupon a lien is placed on it as security for repayment of the debt.

Mortgage Backed Securities (MBS) A security containing a large number (pool) of individual mortgages. When investors purchase one unit of the MBS, they receive a prorated interest in the pool.

Multiple listing service A service generally organized by local real estate brokers, which gathers all of the local property listings into a single place so that purchasers may review all available properties from one source.

Network externalities The quality of a network that its value to each user rises as the total number of users increases.

Real estate Land and the structures built on the land. The structures may be either for residential housing or for commercial uses.

Real estate investment trusts Publicly traded companies that hold a portfolio of real estate properties, similar to a mutual fund, but which hold properties, not common stocks.

Residential real estate Real estate, which is used for housing, including single family, multifamily (apartments), condominium, and cooperative formats.

Transactions costs, real estate The costs of carrying out a real estate transaction, including fees for appraisers, brokers, and other intermediaries.

Virtual community and virtual network A group of individuals with a common association or activity but connected only through the Internet.

Virtual tours An Internet method that provides the visual experience of "walking" through a property.

CROSS REFERENCES

See *Internet Literacy; Online Banking and Beyond: Internet-Related Offerings from U.S. Banks; Online Communities.*

REFERENCES

Bardhan, A. D., Jaffee, D., & Kroll, C. (2000). *The Internet, e-commerce and the real estate industry* (Research Report). Berkeley, CA: Fisher Center for Real Estate & Urban Economics, University of California, Berkeley.

Diaz, J. III. (1997). An investigation into the impact of previous expert value estimates on appraisal judgment. *Journal of Real Estate Research, 13*(1), 57–66.

Dotzour, M., Moorhead, E., & Winkler, D. (1998). The impact of auctions on residential sales prices in New Zealand. *Journal of Real Estate Research, 16*(1), 57–72.

Muhanna, W. A. (2000). E-commerce in the real estate brokerage industry. *Journal of Real Estate Practice and Education, 3*(1), 1–16.

Real Estate Broker's Insider (1998, February). Retrieved February 2, 2002, from http://www.brokersinsider.com/

Schlauch, A. J., & Laposa, S. (2001). E-tailing and Internet-related real estate cost savings: A comparative analysis of e-tailers and retailers. *Journal of Real Estate Research, 21*(1/2), 43–54.

Shapiro, C. & Varian, H. R. (1999). *Information rules: A strategic guide to the network economy*. Boston: Harvard Business School Press.

Mortgage Bankers Association of America (2001, May 8). *Consumers use Internet in mortgage, homebuying process* (Press Release). Retrieved February 2, 2002, from http://www.mbaa.org/news/index.html

Worzala, E. M., & McCarthy, A. M. (2001). Landlords, tenants and e-commerce: Will the retail industry change significantly? *Journal of Real Estate Portfolio Management, 7*(2), 89–97.

FURTHER READING

Georgia Institute of Technology, College of Computing, Graphics Visualization and Usability Center, http://www.gvu.gatech.edu/user_surveys/

http://www.pikenet.com

http://www.realtor.com

U.S. Department of Commerce, E-Commerce Section, http://www.doc.gov/Electronic_Commerce/

Research on the Internet

Paul S. Piper, *Western Washington University*

INTRODUCTION

Whether one is searching for inexpensive airline tickets, or looking for information on corpus collosum bisection, the Internet is an invaluable resource. Studies indicate that students, both graduate and undergraduate, are using the Internet for research in unprecedented numbers and in some disciplines more frequently than any conventional resources (Davis & Cohen, 2001; Davis, 2002). Faculty, scholars and other researchers are not far behind (Zhang, 2001). The Nintendo generation is consistently more acculturated to using online resources than previous generations, and one can expect this trend to continue into the foreseeable future. How did we get here?

In July of 1945 Vannevar Bush, Director of the Office of Scientific Research and Development, published an article in *The Atlantic* entitled "As We May Think" (Bush, 1945). In this article, Bush envisioned a research tool that would link researchers globally, place the contents of libraries at a researcher's fingertips, and monitor what pathways of research each researcher utilized, so that pathways could be easily replicated. The research tool was dubbed the Memex. An interface of software, system and hardware, the Memex was conceptualized exclusively to accelerate and enhance research. Today, the Internet is rapidly actualizing the conceptual Memex. Within the past 20 years the Internet has radically altered the way research is conducted in nearly every area of academia, commerce, and society.

The Internet was initially created to enable computers at universities and government facilities to share files, thus enabling computer scientists and researchers to exchange and disseminate data and information over great distances. The initial experiment, dependent on the creation of packet switching, by the ARPAnet project, achieved initial success on Labor Day weekend, 1969. Since then, the primary Internet developments have been rapid and, with the exception of electronic mail (e-mail) and entertainment applications, primarily research driven.

By 1971 there were 23 hosts connected: all universities, government research facilities, developmental companies (such as BBN), and independent research facilities. By 1973, the Internet was international, with a host computer at the University College of London, England. From there on, the growth has been quite literally explosive.

The development of research enabling and enhancing software for the Internet has followed suit. In 1971 Ray Tomlinson invented a piece of software that may represent the most common Internet use today—e-mail. Within months, Lawrence Roberts, the chief scientist at ARPA, wrote the first e-mail management program, enabling the development of electronic lists (such as LISTSERV), which have become a critical component of scholarly communication and research. 1974 heralded the creation of Telenet, the first commercial packet data service, which opened the door for file transfer between any computers on the network.

Early Internet researchers, primarily academics, used these applications to share and disseminate information, data sets, and manuscripts. Unlike snail (postal) mail or fax, information could be instantly disseminated to all the members of an online community of scholars, and

feedback was potentially instantaneous. Also unlike fax or snail mail, data were instantly computer useable.

Virtual space arose in 1979 with the creation of multiuser dialog software, originally created for gaming interests, and USENET. USENET and bulletin board systems (BBSs) represented a quantum leap in Internet popularity and information sharing. The 1980s saw the growth of networks dedicated to various research groupings, such as NFSNET and BITNET. Internet relay chat was developed in 1988.

Until HYTELNET (the first Internet directory) and Archie (the first Internet search engine) were released in 1989, researchers were dependent on their own haphazard exploration, the word of colleagues, and information posted on mailing lists and newsgroups to identify Internet content. Networks, and network nodes (computers on the network), were extremely limited and regulated. There was little or no open access for outside users. Furthermore, users needed to be relatively skillful at using the UNIX operating language to extract content from the Internet. Archie changed forever the way content was accessed on the Internet.

The advent of Gopher software in 1991, and its search software, Veronica and Jughead, in 1992, completed the conceptual shift of the Internet from a series of nodes on a network to a collection of databases, virtual libraries, and document delivery systems. Searches could be executed and relevant information returned, printed, or downloaded. Search results were low-cost, increasingly global, and instantaneous. In 1992, the number of Internet hosts increased to exceed 1,000,000 in 40 countries.

At this time there was little commercial content or traffic on the Internet, and it was essentially a playground for online searchers and researchers. The Internet Hunt, an online game where a player located the answers to questions in the most efficient way, gained enormous popularity. The Internet was inadvertently creating a culture of searchers.

The World Wide Web (WWW), created in 1992, remained relatively obscure until the creation of the Mosaic browser in 1993. During this year, the growth of Gopher was 997%; the growth of the WWW was 341,634% (Zakon, 2001). While the Lynx browser (a graphicless browser) enjoyed brief success, and is still used faithfully by some, it has for all practical purposes disappeared. Netscape followed on the heels of Mosaic, and researchers have seldom looked back.

On a different front, Project Gutenberg, begun in 1971 by Michael Hart at the University of Illinois, began hosting entire books (classics and older texts) in the public domain in ASCII text format. These texts could be read and searched online, or downloaded. Hart's philosophy was based not on that of a lending library, but on what he termed "replicator technology." Replicator technology advanced the theory that every digital file possessed infinite possibilities for replication. Project Gutenberg also launched the phrase "e-texts." Gutenberg, and similar projects, such as Net Library, has enormous value for researchers and scholars because they allow full-text online searching of manuscripts.

Within academic research, academic journals are the most critical medium for reporting and tracking research. The Internet offered not only another medium for dissemination, but also an opportunity to incorporate other e-text features, such as hyperlinks, into the journals. By 1994 there were over 40 electronic journals on the Internet (Roes, 1995).

In 1994 the first true WWW directory, Yahoo!, was created by David Filo and Jerry Yang at Stanford University. During the same year, the first WWW search engine, Lycos, was invented at Carnegie-Mellon University. There is no way the import of these two inventions can be overestimated.

As the Internet has grown to several billion pages, so has its usability as a research tool. Of particular note is the Internet's value to developing nations (Warren, 1995). Information access and storage is problematic in developing nations, with the adverse factors of climate, space, economics, and censorship impacting text-based collections. The Internet, with its plethora of free information, is viewed as a literal godsend, allowing information to flow relatively unfettered from the information rich to the information poor.

DIRECTORIES

Internet directories represent an attempt to organize Internet content, and are invaluable for conducting research on the Internet. While directories are often referred to as search engines, they are an entirely separate endeavor, and they often incorporate search engine software in their enterprise. Directories range from a simple list of hypertext links related by subject, to an extensive and highly detailed taxonomic organization of subject categories such as The Open Directory Project.

While directories vary greatly as to design, scope, and purpose, there are several qualities that they share. Directories are typically created by people, so there is at least some degree of human judgment that influences the selection of Internet sites included. This gives them, potentially, an inherent authority, but in some cases can also bias them. Directories are organized according to particular, and often unique, schema. And directories often feature site descriptions, annotations, and/or evaluations. On the negative side, directories are limited, sometimes severely, in scope and content. Since sites are typically reviewed by people, and annotations or brief descriptions created, directories are labor-intensive, and this naturally limits their size. For the same reasons directories are extremely difficult to keep up-to-date. Given these constraints, however, directories created by academics or other subject authorities are a critical resource for researchers.

The first online, hypertext directory on the Internet was HYTELNET, a worldwide directory of telnet sites on the Internet compiled in 1991 by Peter Scott of the University of Saskatchewan. The directory was multidimensional, and featured library catalogs, databases, bibliographies, BBSs, government sites, and a number of other resources. Its major flaw was the alphabetical organization, which created lengthy, unwieldy lists. Furthermore, content was often impossible to deduce from the nomenclature. These faults aside, however, HYTELNET represented a major breakthrough in accessing Internet information. One of its greatest assets was the list of links to library catalogs

from every continent. Complete with log-in instructions, this access was the first time researchers could access libraries across time and space and examine holdings.

The next major development in directories was the creation of Yahoo!. Yahoo! was conceived as a personal map of the Internet, highlighting sites useful to Filo and Yang's Ph.D. research, as well as Web sites of personal interest. The magnitude of content eventually necessitated a division into categories, and subsequently subcategories. Yahoo! celebrated its first million-hit day in the fall of 1994, and a new era of locating content on the WWW had successfully germinated.

The number of Internet directories has continued to multiply, even as their scope divides. Directories with vertical subject emphasis have emerged, often hosted by academic societies, libraries, universities and research facilities. These directories are extremely valuable for research purposes. As directories propagated, metadirectories, or directories of directories, materialized. These are exceedingly useful tools for locating directories that index narrow and obscure subject areas. A search for a subject combined with the phrase "Internet directory" (for example, herpetology and "Internet directory") in a good search engine will often yield valuable results.

While most large, general directories such as Yahoo! or The Open Directory Project represent the compilation of numerous contributors, there are a great number of personal directories that are exceedingly useful for researchers. These vary in size from a short list of related links to an extensive directory such as Alan Liu's Voice of the Shuttle, a humanities directory hosted by the University of California, Santa Barbara.

Rings

Rings, although a different entity, share some characteristics with Internet directories. Rings exploit the interconnectivity of the Web by linking sites within a particular specialization. Rings consist of member sites which all have a stake in the authority and credibility of the entire Ring.

While the creation and maintenance of early rings, such as EUROPa, were left to the participants, commercial software and servers now offer ring creators and participants the luxury of HTML templates, scripting language, and automatic updates. Some of these commercial services, such as Yahoo! Rings and WebRing, offer directories of the thousands of Rings using their software or servers. Although rings are particularly popular in the areas of sports, recreation, entertainment, and other areas of popular culture, the serious researcher should not neglect this resource, since they often contain sites that are otherwise unavailable. Some Ring sites require registration to use, while others allow casual users to browse freely.

Weblogs or Blogs

Web logs, or blogs, are a recent hybridization of personal directories, electronic newsletters, clipping services, and alert services. Blogs are collections of links to Internet information in specific areas that the author feels are useful and important. Typically authored by one person, blogs vary highly in content, scope, and purpose. They are often idiosyncratic and represent strong opinions and passions. They are updated frequently and often feature brief annotations. Some blogs offer search capabilities. Directories of blogs exist on many of the major Internet directories.

SEARCH ENGINES

Before a user can effectively utilize Internet information, he or she needs to know that the information exists. The most widely used tools for locating information on the Internet are commonly called search engines. The name is somewhat of a misnomer, however. What Internet users commonly refer to as a search engine is actually an information retrieval system (Liddy, 2001). A search engine typically consists of several components: a document locator, a document processor, a query processor, a search and match processor, and a relevancy ranking delivery system.

Pre-WWW Search Engines

The pre-WWW search engines were rudimentary products based on early database applications, but unlike database software, they were dealing with a chaotic, often disconnected collection of data, information, and files in varied formats. The primary problems for researchers with the rudimentary search tools, Archie, Jughead, and Veronica (which searched Gophers), and wide area information servers (WAIS) were three-fold. File names or Gopher menu headings, the primary search field, often had nothing to do with content. There were many discrete networks and no one tool searched all. And each had different commands and handled information differently. WAIS deserves special mention because it was the first full-text Internet search tool.

Gopher was fated by the fact that its creation was virtually simultaneous to the creation and release of the World Wide Web by Tim Berners-Lee at CERN in 1991. While the WWW floundered until the creation of a decent browser, Mosaic in 1993, use of the Web grew 341,634% and Gopher 997% over the next year, and Gopher, Veronica, and Jughead faded into relative obscurity.

WWW Search Engines

With the advent of the WWW and capable Web browsers, Internet search engines began to actualize the capabilities of what researchers had come to expect when using commercial database services. The evolution of these information retrieval systems has been extremely vigorous.

The primary characteristic of WWW search engines is that they typically search the full text of the documents they have indexed. Full text searching had existed on the Internet before these applications but it suddenly became the status quo. This represented a departure in technique for many researchers, who were accustomed to authority language such as descriptors and subject headings.

Web search engines obtain their index content in several ways, but the most prominent is by using bots, spiders, crawlers, or other intelligent agents. These automated agents utilize hyperlinks to "travel" the expanse of the Web, and identify and index Web content. Most search engine providers have numerous crawlers working

in concert. Providers also rely on content that is submitted by content creators. Even today, none of the most sophisticated Web search engines come close to indexing all Web content. There are several reasons for this that will be discussed later in this section. Once the content is identified, the document processing component identifies the elements that require indexing, removes stopwords, performs stemming (identifies the root of a word for truncation searching), and stores the indexed terms in an inverted file. The program then weights the terms, typically by weighting the term's position within the document's fields, the number of occurrences of the term, and other components. This component differs between search engines, and leads, in part, to different result lists and/or differing position of results when the same search is submitted to several search engines. Different search engines also process varying amounts of information from each site, with some concentrating on the first few pages, others indexing the entire documents. This is why it is critical for serious researchers to use more than one search engine. When all these processes have been completed, the results are stored in an index, or searchable database.

Components of a successful search consist of the user's query, and how that query is constructed and processed. Competent researchers successfully strategize their search by teasing out key words, synonyms, and names and exploiting the advanced features of the search software (fields, truncation, domain, and phrase searching), which differ from engine to engine.

Internet search engines are unique with regard to what search options they offer users and how they process a search query. Search options include using natural language, Boolean operators, truncation, proximity operators; and/or searching by phrase, field, file type, language, case, URL, or "links to URL." They typically do not include authority vocabulary (descriptors and subject headings), although some software allows for metatag field searches. The lack of consistent authority language is still a major obstacle to increasing relevance, but the range of content and content providers on the Internet makes authority control virtually impossible.

Search engines also vary considerably in how they create relevancy rankings for search results. The most common methods include how often the term or terms occur; in what proximity they occur; in what fields they occur; and how many links exist to a Web page. In addition, it is becoming more common to charge money for search return placement. This procedure, called "paid placement," guarantees those willing to pay placement within the top results. Paid placement is seen as a detriment to fair relevancy, but typically involves commercial (dot-com) sites which often do not factor highly in research. The more upright search engines either segregate paid results or do not accept paid placement.

How results are displayed also varies considerably, with relevancy-ranked annotated lists the most common search outcome. Some newer search engines employ artificial intelligence programs to process requests and display results as related topics, subcategories, and alternative topics (for example, a search for "Dolphins" might return the categories "Marine Mammals," "Whales and Dolphins," and "Miami Dolphins"). Users can then modify their original search based on the most accurate category. While search engines of this nature are designed primarily for novice searchers, the categories and concepts returned can prove valuable by offering the researcher additional or related terminology.

General search engines are considered horizontal search tools, in that they search across all subject areas. Specialized search engines are vertical search tools and address a researcher's need to explore in depth a particular subject area. They have evolved rapidly and exist for numerous subjects. A search on a general search engine for specialized engines, for example ("search engine" geology), will generally give researchers the tool they desire, as will directories of search engines such as Beaucoup or Search Engine Colossus.

Meta-search engines are search engines that have no database or inverted file of their own, but run a query simultaneously against a set of independent Internet search engines. Some meta-search engines allow the user to configure which engines are searched, while others operate with default settings. These tools can prove extremely useful for obscure topics because they cover so much ground. Their major drawbacks are that meta-search engines cannot handle customized searches well, due to the vagaries of how individual engines handle a search, and some of them fail to remove duplicates. Their software continues to improve, however, and some of the new generation engines are proving extremely useful for a wide range of needs.

In addition to search engines, a number of commercial products exist which add value to results in terms of organization, categorization, storage, and note integration.

Virtually all fee-based professional journal indexes and databases now have Web interfaces. These can often be readily accessed through local libraries. Many of these databases index Web sites as well as journal articles. Versions of government-funded databases such as Medline, Agricola, and ERIC are free of charge to all users and offer excellent user interfaces. There are also some terrific databases such as Ingenta and Scirus that offer document delivery options at a cost, but offer searching and bibliographic information free of charge.

Ask an Expert Services

These are typically commercial services that are modeled after a library reference service, but actually have their precursors in BBSs such as EasyNet. A user is connected with an "expert" in the chosen field, who then provides research assistance. While these services are generally free, some do cost. And while they can prove useful, their primary drawback is their lack of screening "experts."

Maximizing Search Engine Effectiveness

The difference in all areas of search engine performance is truly staggering. For this reason, an effective researcher needs to consistently use and master several search engines, exploiting their differences. It is critical to explore help pages, use advanced search options, and practice repeatedly. Keeping current on search engine developments is also critical, since this technology changes daily. For

this purpose, academic and trade journals, newsletters, and various online publications dedicated to this technology are invaluable.

Document Code

Advances in HTML code, such as extensible markup language (XML), enhance machine-readable structure within the document itself, thus augmenting search precision. XML, an extension of SGML, is a structural ordering system which defines a range of discreet document sections. XML enables generic SGML to be served, received, and processed on the Web similar to HTML, thus making available a number of discrete document sections or fields that are not currently available in standard HTML. The potential growth of XML holds great promise for researchers dissatisfied with the lack of precision in the structure of standard HTML documents.

The Invisible Web

The Invisible Web refers to information on the Internet that has value but is not typically indexed by search engines. There are two general categories of material that go undetected.

The first category consists of information not detected by the various crawlers and spiders search engines use to index Internet content. This lack of detection is occasionally due to Web pages that possess no external hyperlinks. Pages of this type are considered disconnected. If a crawler cannot get into the content, it cannot index it, and since most crawlers rely on hyperlinks as their conduit, if there are no links in, crawlers cannot find them.

Site depth also accounts for invisibility. Some search engines do not index entire documents, but specify a certain percentage or number of pages they do index. Others only index pages to a certain level, which means that information several levels below the home page may not be indexed. Information on site coverage is characteristically found in the "About" information on a search engine site, and it is well worth investigating when evaluating search engines for research potential.

Further information can remain hidden from search engines because it contains content that cannot be interpreted by a crawler. Among the thousands of file types on the Web, there is only a handful that crawlers capably identify. Some crawlers are more adept than others, and a few search engines allow the user to specify file type, retrieving file types with extensions such as doc, pdf, ppt, gif, jpg, mpg, aud, and others.

Other obstructions to crawler technology are online forms. Forms act like gates that need to be entered before content becomes available. Crawler technology is coevolving and will soon be capable of passing through at least some forms.

Dynamically generated pages are yet another impediment for crawlers, since these are pages created by scripting language in response to a database query, form, or other user interaction. These vary dependent on the user request, and it is impossible to index them all. The best a crawler can do is index the interface gate, which may be adequate if enough terminology is provided by the page author.

There are also many specialized databases of journal articles, reports, newsletters, corporate and organizational information, and so forth, whose content is hidden from crawlers. To search these databases effectively, one needs to identify the database and search it independently. An Internet search engine will not return Medline or ERIC results. These remain invisible to crawlers.

The other primary category of invisible information is hidden information. Hidden information is not detected by crawlers because the author or webmaster desires it so. Webmasters use different methods of blocking access to sites (Sherman & Price, 2001). Blocking protocols, such as the robots exclusion protocol (which creates a list on the server of files not to be crawled), or "noindex" metatags will keep many crawlers out. Passwords are a more effective method of thwarting crawlers and keeping information private.

INTERNET COMMUNICATION

Electronic Mail

Electronic mail was originally created as a convenience or entertainment but quickly became a vital component of the research process, giving researcher's global access to people and information instantly. A major breakthrough occurred when it was discovered that the transmission of documents, graphics, presentations, and other files was possible by attaching them to e-mail messages. This quickly became an invaluable tool for researchers who were jointly writing and editing. In addition, e-mail grew in popularity as an alert service for articles, publications, products, conferences, and other events of interest to scholars and researchers.

Mailing Lists and Newsgroups

An evolution of e-mail that has become de rigueur for many researchers is that of mailing lists and newsgroups. Mailing lists employ software that disseminates or makes available information, commentaries, and questions simultaneously to all subscribers of the list. The first mail list designed specifically for research, THEORYNET, originated at the University of Wisconsin in 1977. In these, and all other mailing lists on ARPAnet and the early BITNET, human intervention was required to add subscribers to the list and distribute the e-mail. By 1985, BITNET had replaced ARPAnet as the academic and research network. Their mail list, called LISTSERV, was also person-moderated, and experienced enormous delays in both subscriptions and mail delivery.

LISTSERV software, invented in 1986, automated this process. Through automation mailing lists have flourished. Catalist, the primary directory of LISTSERVs, listed 210,949 at the time of this writing. While the term LISTSERV has become genericized to represent any similar software or group, there are numerous other push mailing list technologies, such as (Majordomo or Gnu) in existence, and other directories that index them.

LISTSERVs can be moderated or unmoderated. Moderated lists are mailing lists that employ a human editor to filter submitted messages before they are distributed. Unmoderated lists disseminate to all members without

human intervention, although most software is now capable of some filtering. Unmoderated mailing lists tend to be larger and function more quickly. LISTSERVs can also be public or private. Public lists have open enrollment, while private lists require members to meet specific membership criteria, such as enrollment in a professional organization, employment or enrollment at a specific institution, or designated academic or professional status.

Newsgroups are similar to mailing lists in that they sort and distribute mail according to subject specialties, but they utilize pull technology, which requires users to access a server to read messages, rather than receive them as e-mail. Newsgroups generally thread, or group their messages, into topics, giving users the option to choose which discussions, within a general topic, to interact with. Newsgroups have become a major component of online instruction.

USENET is the largest news group on the Internet to date. USENET was originally created in 1979 by graduate computer science students at Duke University and the University of North Carolina as an alternative to the ARPAnet, which many students and researchers contended was too restrictive. USENET evolved into hundreds of subject groups, arranged hierarchically. Examples are alt (alternative), humanities (fine art, literature, and philosophy), sci (the sciences), and soc (sociology and social concerns). USENET groups evolved into one of the most vibrant areas of intellectual, cultural, and recreational discourse on the Internet. They have recently been subsumed by a commercial search engine, but remain fully functional, and searchable, including the archives. They contain a plethora of extremely unique information, but some care should be exercised since identity and informational claims are not verified. Even with these limitations, USENET represents one of the most underused research tools on the Internet. There are currently many other mail and newsgroups, many of them managed by commercial interests.

Bulletin Board Systems

Bulletin board systems are computer systems used as information centers and forums for particular interest groups or locations, and as such had limited, though unique usability. BBSs typically required individual dial-up access and faded out as the Internet developed. It could be argued that portals represent an analogous conceptual approach.

Chat Rooms and MUDS

Multiuser dungeons (MUDs) were inaugurated at England's Essex University in 1979 for the purpose of allowing multiple users to create imaginary worlds and play fantasy games. Along with their successor MOOs (object-oriented MUDs), these virtual spaces have enormous potential as brainstorming and discussion venues. They are currently underused for this purpose.

Online Conferences

Online, or virtual conferences, are online versions of a physical conference, capitalizing on the unique features of a Web environment. Online conferences may utilize a combination of virtual space (MUDs and MOOs) and other real-time chat technologies; online video and audio feed; white board technology; e-mail and mailing lists; online reports, articles, and supplementary documents; and related hyperlinks. Since virtual conferences typically model academic conferences, virtual conference attendees register, often paying money, choose a conference track or tracks, and read, listen to, or watch the conference presentations and any supplemental material provided. The track participants typically join a mail list or partake in chat sessions, which mirror conference discussion and networking, a key element of conferences. Presenters are generally available via e-mail for questions and discussion as well. Conference products include bibliographies and proceedings, as well as electronic updates.

While there are numerous criticisms of online conferences, there are many advantages. There is no travel involved, thus allowing underprivileged participants and those with time constraints to participate; the cost is minimal; materials can be read and presentations attended in a flexible manner and over long periods of time; hyperlink technology provides a plethora of supplemental materials; and e-mail connectivity allows interactivity with participants and presenters, many of whom are essentially off limits at nonvirtual conferences.

LIBRARIES AND THE INTERNET

Providing research access and assistance is one of the essential functions of libraries. With the advent of the Internet, there has been much discussion as to how useful libraries remain. The fact is, however, that libraries have embraced the Internet since its origin and are reliant on it for operation. They are extremely proactive in the development of research-based Internet applications and services, and remain critical in assisting in almost all aspects of Internet use and evaluation.

Academic libraries, often affiliated with the early ARPAnet partners, established an early presence on the Internet by hosting their catalogs online. Since these academic libraries were among the most prestigious in the world, this sudden, global sharing of collections enabled researchers to locate rare or previously unknown materials and expanded the scope and depth of their research enormously. As the Internet developed with the WWW, libraries migrated to the Web as the primary medium for catalog delivery and other services.

The technological development of online catalogs over the Internet led to cooperative or union catalogs. MELVYL, a union catalog comprising all the University of California schools, began development in 1977. HYTELNET was the first Internet software to provide access to global library catalogs. Libdex is a contemporary Web-based directory and search engine for libraries on the WWW.

Libraries without Walls

The concept that the materials contained within the physical structure of a library no longer limited the library's collection first surfaced in the mid-1980s. Access to library

catalogs from anywhere in the world accelerated the development of interlibrary loan, or the ability to borrow materials from other libraries. Access was suddenly on equal footing with ownership. Electronic access in concert with interlibrary loan expanded the concept of the library far beyond its physical space.

Online Computer Library Center, Research Libraries Information Network, and other similar services provide Internet networks that share copy cataloging, allowing catalog updates instantly on a global scale, and global access to collections, thus greatly accelerating the speed and precision of interlibrary loan.

Current access to databases, directories, full-text books and journals, and other digital collections has continued to expand the concept of the library without walls.

Librarians and the Internet

It is safe to say that librarians continue to perceive the Internet as a double-edged sword. The primary complaints are a lack of content quality and quality control, a lack of organization, and the labile nature of much Internet material, which frustrates attempts to catalog and retrieve it. Another fear, often unstated, is that the overwhelming popularity of the Internet could, at least partially, replace libraries and librarians. On the other hand, libraries embrace the Internet as a medium for hosting collections and databases; for exchanging books and other materials; for creating authoritative directories and portals; for doing reference; and for a host of other uses. Librarians are also key teachers of Internet literacy. Rather than diminishing the need, the Internet has in reality created an increased need for librarians and other information professionals.

Internet Reference Service

The ability to answer reference queries at a distance has always been an important component of library service and has proved invaluable to research. Distance reference queries were initially communicated by letter, telephone, and fax. E-mail gives librarians a tool for answering queries and/or sending research strategies without constraints of time and locale. This service has become instrumental for distance learning.

A new development in e-mail reference service is the development of consortia that share and disseminate requests, often automatically. Internet users submit a request form whose databased fields are linked to another database of subject experts. Based on the topic, the request is sent to a librarian (who may reside in an entirely different locale) with expertise in that field. The most comprehensive example of this service at this time is the Collaborative Digital Reference Service hosted by the Library of Congress and piloted in the year 2000. This service is global and provides professional reference service to researchers 24 hours a day, via an international digital network of libraries and related institutions.

Some libraries have experimented with real-time interactive video reference over the WWW. Dependent upon compatible systems and webcams, this technology has not developed quickly, but there is optimism for future applications.

Cataloging the WWW

Librarians and professional researchers, as well as commercial enterprises, have been struggling with methods of cataloging the Internet. In many ways the task is fruitless, since millions of pages of information are added daily, and a significant fraction of those disappear or change location. Still, this has not curtailed efforts.

There are two primary approaches in the library world. The first is selective cataloging, and this is commonly practiced by libraries of all type. It consists of cataloging Internet information that fits selection criteria, such as quality, authority, longevity, uniqueness, and so forth. Information of this type is identified by individual bibliographers and selectors, as well as cooperatively. The second approach is to somehow identify Internet resources by subject headings or other authority languages.

Online Cooperative Library Center (OCLC) is exploring both of these approaches. The Dublin Core/Metadata Initiative calls for the development of interoperable online metadata standards, analogous to Library of Congress subject headings that support a broad range of purposes and business models. The resulting descriptors would be inserted into the metatag field of a Web document, hopefully yielding a much higher degree of search precision.

While the Dublin Core project addresses code-based content, OCLC is also mainstreaming a project called CORC. CORC is an attempt to create a shared database of librarian-selected and cataloged Internet sites. CORC uses an on-the-fly cataloging process, pulling data automatically from HTML fields. Membership in the CORC collaborative enables the member libraries to access all CORC Internet records. CORC members will have access to an ever-increasing number of stable, superior Web documents and information. Users of these libraries will reap the benefits.

ELECTRONIC JOURNALS

The linchpin of academic research has, and continues to be, academic journals, although there are signs this foundation is eroding slightly due to economic concerns. In addition, traditional academic journals are losing their current awareness function to mailing lists, electronic journals, electronic alert services, and other Internet-based services.

What we currently know as electronic journals evolved out of projects such as ADONIS, which replicated text-based journals by placing article images on CD-ROM. Full-text databases on CD-ROM were also precursors. The first electronic journals on the Internet began appearing in the late 1980s, and soon after several directories appeared. The largest of these was sponsored by the Association of Research Libraries and featured the LISTSERV NEWJOUR-L as an alert for new additions. Dubbed the Directory of Electronic Journals, Newsletters and Academic Discussion Lists, the directory contained 440 entries by April 1994. These early journals were accessed in a variety of ways including e-mail/USENET, Gopher, FTP, WAIS, and WWW, with ASCII being the primary format (Roes, 1995).

The increasing popularity of the WWW as a medium for accessing and disseminating information has created

a profusion of Web-based electronic journals, as well as other publications such as newsletters, newspapers, magazines, books, and hybrids. While many of these are electronic versions of print versions, some are unique to the Web. Regardless, there are considerable advantages to the Web format which include quicker peer-review; electronic preprints; full-text searching; no article size limitation; quicker delivery time; open access; exploitation of hyperlinks to connect to other related articles, authors, or supplemental material; the inclusion of multimedia and other graphic, auditory, or database file components into an article; links with online discussions, surveys, and other interactive components; and "spaceless" archival storage. While the migration of prestigious academic journals to WWW format has been more cautious than originally anticipated, there are currently very few academic journals that do not offer some type of Web presence. Even many text-only journals are commonly using the Web as a source of supplemental material.

Many specialized electronic journals are now available for table of contents browsing, and in some cases key word searching of article titles and abstracts. Some large publisher aggregates now offer search features similar to journal databases, free of charge.

EVALUATION OF INTERNET CONTENT

Faulty information on the Internet is, and will plausibly always be, a major problem. The fact that nearly anyone can publish on the Internet creates both an environment of exhilarating freedom, and one that lacks any overall quality control. The filtering and editing traditionally furnished by content providers is now largely in the hands of the Internet user. Superimposed on the particularities of individual sites is a major structural misconception. In traditional venues for the exchange of information, there is a frame or package that reflects content. This frame creates reader or viewer expectations of content. A popular magazine, for example, creates the expectation of content that differs radically from that of an academic journal. The same can be said of newspapers, radio, and television. For example, some radio talk shows are liberal, others conservative; some television stations feature primarily comedy, others old films, and so forth. Readers, listeners, and viewers are either immediately aware, or quickly become aware of these differences. With the Internet, however, the package, for example, a list of results from a search (which is roughly analogous to a table of contents) may contain content that differs fundamentally with regard to quality and orientation. Wise Internet use, or Internet literacy, can be achieved by following some fairly common-sense guidelines.

Authority and Bias

Innate within the information contained in any publication is its authoritative credibility. It is no different on the Internet. Information hosted by the National Library of Medicine or the Louvre has a credibility that may be lacking from a personal Web page. A personal Web page constructed by an acknowledged authority in the field has a credibility that a personal Web page constructed by a high school student does not have. Furthermore, even though we recognize that all information is on some level biased, the degree of bias varies considerably between information hosted by NASA and that hosted by a racist group.

URL Clues

One quick way of evaluating quality is by examining domain names, which, although not fool-proof, can provide clues. Generally speaking, government (dot-gov) sites are accountable to the public for information hosted on their sites, which makes them relatively reliable and unbiased. The information on organizational (dot-org) sites is susceptible to the viewpoint and bias of the organization, which can affect both the veracity and slant of the information. Information hosted by commercial (dot-com) sites is most problematic. Generally speaking, commercial sites are out to sell a product and will not host information that undermines that purpose. It is recommended that one obtain corresponding verification if using facts and figures obtained from commercial sites. The primary exceptions are fee-based sites such as indexes and databases and electronic newspapers, magazines, journals, and books, which are generally as reliable as their print counterparts. Sites bearing the dot-edu or educational domain are fairly reliable, but personal philosophies typically go unedited. One can generally find solid, credible information from departmental or research-oriented pages.

There is growing confusion and blurring of domain names, particularly dot-org and dot-com, with a number of for-profit endeavors opting for dot-org domains. In addition a number of new domains (dot-name, dot-museum, and dot-info) are appearing, which in addition to country domain codes muddy the water considerably. Still, the domain name is a good starting place.

Another URL clue is the tilde (~), which often signifies a personal page hosted by an Internet service provider or a company, organization, or institution. As such, the authority and bias of the host is not applicable, since personal sites are not always reviewed for content by the hosting organization. When critiquing the quality of a personal page, reflect first on the standing and authority of the page's author.

URLs can be checked at a variety of domain lookup sites to verify author, contact information, and so forth, and this can provide users with important clues as to the intention of the site's creator. I suspect that site author names and site names can be searched on search engines and journal or newspaper databases, occasionally turning up exposes and countering or verifying information. USENET searches can also turn up interesting results.

Audience

The bias and quality of information correlates to the intended audience of the information. Information that preaches to the saved is not the same as information that seeks new converts. Serious Internet researchers, unless they are investigating popular culture, will typically seek information that is addressed to subject authorities, rather than information addressed to the general public.

Currency

Maintaining a Web presence is an ongoing task, and there are numerous Web sites that went up with a bang three years ago and have not been updated since. The very best sites will tell you up front when they were created and when they were updated. Individual articles, reports, or bits of information may also have dates. The current research is not necessarily the best, but it is necessary to know if it is current or not before that judgment can be made.

Coverage

The extensiveness of coverage is another quality to scrutinize when evaluating an Internet resource. Bias will often be reflected in the error of omission rather than the inclusion of erroneous information. This is often true of commercial (dot-com) and organizational (dot-org) sites that want the user to see it their way. Providing a contrast of perspectives would be counterproductive to their purpose.

Language

Always subject a new or dubious site to a quick semantic analysis. Is the language excessively descriptive, comic, inflammatory, or hyperbolic? Such language tends to represent bias or opinion. In addition, be wary of unsubstantiated claims, for example "*The Harvard Law Review* claims. . ." without identifying information describing volume, issue, date, or page numbers.

Page Structure

There are different philosophies in Web design, and these vary depending on the target audience; hence it is difficult to definitively advance one mode of design over another. However, any design that impedes access is not desirable. Information is only useful if it is available. Excessively busy or chaotic Web sites, sites that are poorly organized, and sites whose design features (color, layout, frames, tables, graphics, animation, Flash, and other multimedia devices) interfere with navigating the site and succeed in obfuscating the information they are purportedly trying to promote. There are also page designs that will attempt to lead researchers to certain information and away from clues to intent.

Links

The "About Us" or related links often provide definitive information about the mission, philosophy, perspectives, and goals of a Web site host. Even spoof and spurious sites can often be exposed this way. Examine e-mail links and related links for clues. The quality and destination of links can often inform one as to the quality of a Web site's content. Numerous broken links can indicate the site has not been updated recently or is ill-maintained.

Source Code

Viewing the source code of a Web page is often less revealing than it once was due to the increased use of Perl, Javascript, ASP, and other scripting languages. Web page creators can now, utilizing scripting language, direct spiders to one page, and browsers to another, thus allowing search engines to classify a site by different terms and concepts than the page the user sees. The growing invisibility of the page source will no doubt continue to increase. It can still be extremely insightful to view source code. An examination of the page title, the metatag field, filenames for linked pages and graphics, and comments can often reveal the true intention of a page.

Misinformation

There is copious information on the Internet that is patently false, some of it by blameless error, some of it by spurious intent (Piper, 2000). Spurious information is problematic because the creators of the information desire to deceive, either for monetary gain or support of a political ideology or other personal rationale.

Counterfeit Web sites and/or credentials are the most common manifestations of spurious information. Counterfeit Web sites mirror the appearance of an authentic site but contain false information. A deviation of direct counterfeit is an authentic site that has been cracked (hacked) and legitimate information is replaced with false, or disinformation. The armed conflicts in the Middle East and the former Yugoslavia have germinated hacking of this nature. Investors have also been victimized by spurious Web pages that counterfeit authentic investment news, such as the fake Bloomberg's page touting the acquisition of ParGain in April 1999.

E-mail fraud, which is an electronic extension of mail fraud and targets specific audiences often using Internet mailing lists, is also a major problem. As might be expected, the elderly are often victims, and the content is often health or financially oriented. E-mail and mail list information is particularly subject to fraudulent claims, and identities can be easily counterfeited. If doubted, these should be verified by legitimate sources.

Spoofs

Spoofs are deliberate parodies, often of existing sites or topics. They are often political in nature, but they can be for instructional or recreational purposes as well. While spoof sites vary considerably with regard to professionalism, intent, and scope, they are generally open about their intentions. Their purpose is most often to parody perceived exaggerations or errors in the original information.

The most common way that parody or spoof pages deceive a legitimate researcher is when underlying pages on a spoof site are returned by a search engine without reference to a home page. These decontextualized pages, often featuring only a fake report or article, are lacking the disclaimers that often accompany the home page and can be misleading even to experienced researchers.

Alternative Views

The Internet provides a more direct mirror of social norms, ideas, and prejudices than any previous medium in history (National Council of Civil Liberties, 1999). This results in an unprecedented array of published alternatives

to status quo or mainstream cultural values and "truths." These alternatives can be viewed as fertile terrain for researchers, or they can be viewed as erroneous information. It is often a personal and subjective call. The prudent method of dealing with alternative information is to unearth and attempt to understand the bias or point of view projecting it and compare and contrast it with authorized findings and beliefs.

Help

There are numerous organizations that track spurious information on the Internet. The Computer Incident Advisory Capability, a site hosted by the U.S. federal government, is a good place to start, and there are numerous other organizations such as Snopes that are excellent in tracking Internet frauds and hoaxes.

CONCLUSION

The condition of global research has evolved significantly due to the Internet. There is no area of research that has gone untouched. Researchers can write, edit, and share papers and reports electronically; read and interact with journals, books, other texts, and media online; examine the holdings of millions of libraries, research facilities, museums, and political entities; perform relatively sophisticated searches on billions of documents; and do all of this, and much more, in real time and from nearly anywhere in the world.

Furthermore, the Internet has made available to researchers information that is impossible to obtain elsewhere, material that is accessible without extensive (or any) cost, and material that represents a wider variety of opinion and cultural concern than is obtainable via any other medium. While writing this chapter, the entire Abraham Lincoln Papers' collection was uploaded at the Library of Congress' American Memory site, fully searchable. And this is just one of millions of comparable items that became available during these few short months.

While it is tempting to wax on about future developments for researchers, it is futile to deal in particulars. At least some of the copyright issues which currently block large amounts of information from appearing on the Internet will be settled; bandwidth will increase enabling new formats and faster transmission; interactivity and intelligent interfaces will further evolve; search engines will become "smarter;" information will become standardized and easier to catalog; and the overall amount of information, particularly from developing nations, will continue to increase at astronomical rates. On the other hand, barring any massive global governmental interventions, information on the Internet will remain relatively uncensored and extremely varied, making it crucial for users to become literate in evaluating content.

GLOSSARY

Internet directories Hyperlinked lists of Internet sites organized by subject, typically selected by an information professional, and often annotated and evaluated.

Rings Internet sites that are related by subject and linked together. Rings often require membership, and self-select for quality.

Internet search engines Software that identifies Web content, organizes and indexes it, searches it, and displays results based on relevance.

Invisible Web Those files on the Internet that are not easily detected by search engines.

CROSS REFERENCES

See *Digital Libraries; Internet Literacy; Library Management; Web Search Fundamentals; Web Search Technology.*

REFERENCES

Bush, V. (1945). As we may think. *The Atlantic,* July, 1945, 101–108.

Davis, P. M. (2002). The effect of the Web on undergraduate citation behavior: A 2000 update. *College & Research Libraries, 63*(1), 53–60.

Davis, P. M., & Cohen, S. A. (2001). The effect of the Web on undergraduate citation behavior 1996–1999. *Journal of the American Society for Information Science and Technology, 54*(4), 309–314.

Liddy, E. (2001). How a search engine works. *Searcher, 9*(5), 39–45.

National Council of Civil Liberties. (1999). *Liberating cyberspace: Civil liberties, human rights, and the Internet.* London & Sterling, VA.: Pluto Press.

Peters, T. A. (1991). *The online catalog: A critical examination of public use.* Jefferson, North Carolina & London: McFarland.

Piper, P. S. (2000). Better read that again: Web hoaxes and misinformation. *Searcher, 8*(8), 40–49.

Roes, H. (1995). Electronic journals: A survey of the literature and the net. *Journal of Information Networking, 2*(3), 169–186. Retrieved March 6, 2002, from http://cwis.kub.nl/~dbi/users/roes/articles/ej_join.htm

Sherman, C., & Price, G. (2001). *The invisible Web: Uncovering information sources search engines can't see.* Medford, NJ: Information Today.

Warren, K. S. (1995). From information statics to information dynamics: The developing world has no alternative. *International Health, 9*(2), 367–395.

Zakon, R. (1993). *Hobbe's Internet timeline.* Retrieved March 6, 2002, from http://www.zakon.org/robert/internet/timeline/

Zhang, Y. (2001). Scholarly use of Internet-based electronic resources. *Journal of the American Society for Information Science and Technology, 52*(8), 628–654.

Return on Investment Analysis for E-business Projects

Mark Jeffery, *Northwestern University*

INTRODUCTION

As the late 1990s came to a close, many companies had invested heavily in Internet, e-business, and information technology. As the technology bubble burst in 2000 many executives were asking, "Where is the return on investment?" When capital to invest is scarce new e-business and information technology (IT) projects must show a good return on investment (ROI) in order to be funded. This chapter will give the reader the key concepts necessary to understand and calculate ROI for e-business and IT projects. In addition, the limitations of calculating ROI, best practices for incorporating uncertainty and risk into ROI analysis, and the role ROI plays in synchronizing IT investments with corporate strategy will be discussed.

What is ROI? One conceptual definition is that ROI is a project's net output (cost savings and/or new revenue that results from a project less the total project costs), divided by the project's total inputs (total costs), and expressed as a percentage. The inputs are all of the project costs such as hardware, software, programmers' time, external consultants, and training. Therefore if a project has an ROI of 100%, from this definition the cash benefits out of the project will be twice as great as the original investment. (In the section Review of Basic Finance we will discuss how this definition of ROI, although qualitatively correct, does not accurately include the time value of money, and we will give a more accurate definition based upon internal rate of return [IRR].)

Should a manager invest a company's money in an e-business project if it has a projected ROI of 100%? There are many factors one should consider when making an investment decision. These factors include, but are not limited to those listed below:

The assumptions underlying the costs of the project.

The assumptions underlying the potential benefits.

The ability to measure and quantify the costs and benefits.

The risk that the project will not be completed on time and on budget and will not deliver the expected business benefits.

The strategic context of the firm; that is, does the project fit with the corporate strategy?

The IT context of the project: that is, does the project align with the IT objectives of the firm, and how does it fit within the portfolio of all IT investments made by the firm?

As discussed in the section Review of Basic Finance, the simple definition of ROI given above is not rigorous enough for good investment decision-making. In addition, the assumptions underlying the model and risks associated with the IT project are key drivers of uncertainty in any ROI analysis. Awareness of these uncertainties and the impact of risks on ROI can significantly improve the likelihood of successful investment decisions.

The return on investment for corporate information technology investments has been the subject of considerable research in the last decade. (For reviews, see Brynjolfsson & Hitt, 1998; Dehning & Richardson, 2002; and Strassmann, 1990.) The most recent research suggests that investing in IT does on average produce significant returns (Brynjolfsson & Hitt, 1996). See the next section, The Information Paradox, for a discussion of this research.

Jeffery and Leliveld (2002) surveyed CIOs of the Fortune 1000 and e-Business 500 companies: Of the 130 CIO respondents, 59% reported that their firms regularly calculated the ROI of IT projects prior to making an investment decision, and 45% of respondents reported that ROI was an essential component of the decision-making process. ROI is therefore an important component of

the information technology investment decisions made in many large companies.

However, an interesting observation is that only 25% of companies responding to the survey actually measured the realized ROI after a project was complete. ROI analysis is therefore primarily used to justify an investment decision before the investment is made. Performing post-project analysis provides valuable feedback to the investment decision process to verify the validity of the original ROI analysis, and the feedback improves ROI calculations in the future. Feedback also enables the weeding out of underperforming projects. Full life-cycle ROI analysis translates into better information to make better decisions, which in turn should impact the returns for the total corporate IT portfolio of investments.

The total IT investments made by a firm can be thought of as a portfolio, similar to a financial portfolio of stocks and options. Each IT investment will have a different risk and return (ROI) and, because capital is limited, selecting the optimal portfolio is a challenging management decision for any firm. The methodology for choosing and managing an optimal IT portfolio is called IT portfolio management. This process often includes the use of scorecards so that executive managers can rate projects on multiple dimensions and ultimately rank projects in relative order of importance to the firm. A typical scorecard will include several categories that help quantify the value of a project to the business and the risk of the project. Note that ROI is typically only one category on the scorecard and that several other factors may have equal or greater importance. In the Executive Insights section at the end of this chapter, an example of the IT portfolio management process at Kraft Foods and its score card used to rank e-business and IT projects are discussed.

In the following section we will briefly review the research literature on returns on investment for information technology and the related information paradox. The third section, Review of Basic Finance, is an introduction to the key finance concepts necessary to calculate ROI. Using these concepts, the ROI for a case example is calculated in the section Calculating ROI for an e-Business Project, and a template is given that is applicable to any ROI calculation. Uncertainty in assumptions and risk are important considerations, and the section Uncertainty, Risk, and ROI shows how to include these factors in the ROI analysis. Specific risk factors for e-business projects that may impact the ROI are also discussed. This section shows how sensitivity analysis and Monte Carlo methods can be applied to ROI models; these are two powerful tools for understanding the range of possible ROI outcomes based upon the cost and revenue assumptions and the risks in the project. The last section, Executive Insights, gives some tools for oversight of technology investment decisions—specifically, questions to ask when reviewing an ROI analysis and how ROI fits within an information technology portfolio management framework for optimal IT investment decisions are discussed.

THE INFORMATION PARADOX

The question of how investment in information technology impacts corporate productivity has been debated for almost a decade (for reviews, see Brynjolfsson & Hitt, 1998; Dehning & Richardson, 2002; and Strassmann, 1990). Productivity is defined similarly to ROI in the introduction—it is the amount of output produced per unit of input—and although easy to define, it can be very difficult to measure for a firm (Brynjolfsson & Hitt, 1998). This difficulty in measurement is similar to the challenges of measuring ROI for information technology and e-business projects. The output of a firm should include not just the number of products produced, or the number of software modules completed, but also the value created to customers such as product quality, timeliness, customization, convenience, variety, and other intangibles.

One would expect that the productivity of the overall economy should increase over time, and this is indeed the case for the manufacturing sector, where the outputs are relatively easy to measure—see Figure 1a. This productivity increase is not due to working harder—because

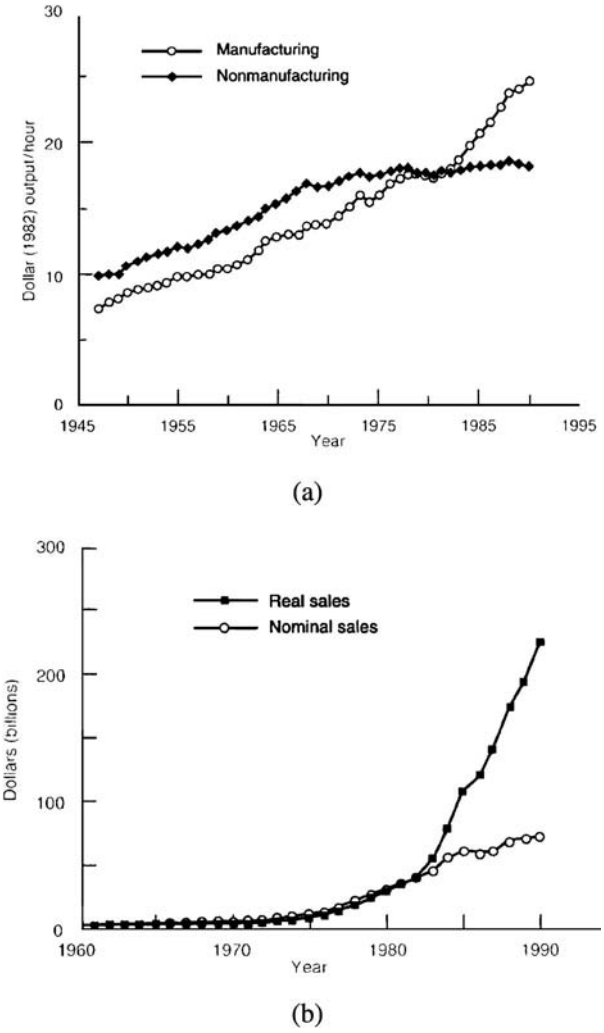


Figure 1: (a) Average productivity for the manufacturing and service sectors. (b) Purchases of computers not including inflation (nominal sales) and sales adjusted for inflation and price deflation due to Moore's law (real sales). The real sales are an indication of the actual computing power purchased. Source: Brynjolfsson (1993). © 1993 ACM, Inc. Reprinted by permission.

although working harder may increase labor output, it also increases labor input. True productivity increases derive from working smarter, and this usually happens by adopting new production techniques and technologies.

The greatest increases in productivity have historically been associated with "general-purpose technologies." Examples are the steam engine and the electric motor. These inventions were applied in a variety of ways to revolutionize production processes. One would expect that computers and the Internet, because they are also general-purpose technologies, should dramatically increase productivity.

However, data in the late 1980s and early 1990s suggested that the average productivity of the U.S. economy in the nonmanufacturing or service sector, which is a primary user of computers and IT, had been constant from 1970 to 1990—see Figure 1a. During this same time frame corporate investments in computers had increased dramatically, so that by 1990 investments in computer hardware averaged 10% of a company's durable equipment purchases. Furthermore, following Moore's law, the number of transistors on a computer chip doubles approximately every 18 months and the speed of computers doubles every 2 years. Hence the "real" computing power purchased by firms increased by more than two orders of magnitude from 1970 to 1990. The apparent inconsistency of IT spending and productivity was termed the *productivity paradox,* and the conventional wisdom of the late 1980s was that there was no correlation between investment in IT and productivity. If the productivity paradox is true, it suggests that firms should not invest in IT because it does not create good ROI.

The problem with this conclusion is that it is based upon aggregate data averages of the entire U.S. economy. These data are averages that measure productivity in terms of the number of products produced. So as long as the number of products increases for the same level of input, the productivity increases. For computers, this accounting works well if they are used to cut costs, but it does not work if they are used to transform business processes or create intangible value. Brynjolfson and Hitt (1998) use the example of the automated teller machine (ATM) and the banking industry. ATMs reduce the number of checks banks process, so by some measures, investing in ATM IT infrastructure actually decreases productivity. The increase in convenience of ATMs goes unaccounted for in traditional productivity metrics. For managers, IT can look like a bad investment when they can easily calculate the costs of the IT investments but have difficulty quantifying the benefits.

In the mid- to late 1990s several research studies were undertaken on new data sets that included individual data on thousands of companies (see, for example, Brynjolfsson & Hitt, 1996; Dewan & Min, 1997; and Malone, 1997). These data enabled researchers to find a significantly better way to measure firm performance. Across all of these research studies there is a consistent finding that IT has a positive and significant impact on firm output, contradicting the productivity paradox. However, these studies also show that there is a significant variation in the magnitude of this payoff among firms.

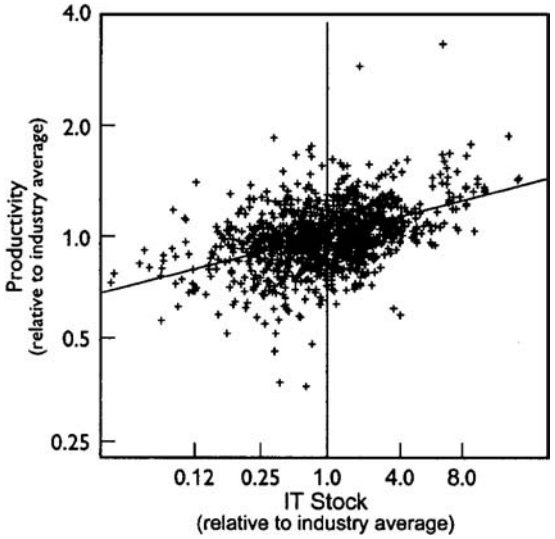


Figure 2: Productivity as a function of IT Stock (total firm IT related expenditures) for a sample of 1,300 individual firms. Source: Brynjolfsson and Hitt (1998). © 1998 ACM, Inc. Reprinted by permission.

Figure 2 is a plot of the variation in productivity and IT investments across 1,300 firms (Brynjolfsson & Hitt, 1998). The horizontal axis (labeled "IT Stock") is the total IT inputs of the firm. The vertical axis is the productivity, defined as the firm outputs divided by a weighted sum of the inputs. Both productivity and IT input are centered at the industry average. The best-fit line is clearly upward-sloping, indicating the positive correlation between IT spending and productivity at the firm level. However, the striking feature of these data is the wide variation of returns. Some companies spend more than the industry average on IT and have less productivity, whereas others spend less and have greater productivity.

The large variations in returns on IT are well known by many corporate executives. For every amazing IT success story such as Dell, Cisco, or WalMart there are many failed or out-of-control IT projects (Davenport, 1998). As examples of these failures, a Gartner survey of executives found that 55% of customer relationship management (CRM) projects do not produce results, and a Bain consulting survey of 451 senior executives found that one in five reported that the CRM system not only failed to deliver profitable growth but also actually damaged longstanding customer relationships (Rigby, Reichfeld, & Schefter, 2002).

The wide variation of returns in Figure 2 is indicative of the fact that there is more to productivity than just investment in information technology. Other factors are just as important—the ability of the firm to exploit organizational change and how the IT investment fits in the context of the firm's strategy in a given industry. Research suggests that there is on average a time lag, of order one to three years, before the benefits of a large IT investment significantly impact a firm's productivity (Brynjolfsson & Hitt, 1998).

In summary, research studies of the late 1980s and early 1990s suggested that there was no correlation between IT investments and firm productivity; this was

called the *information paradox*. However, studies in the mid-1990s based upon firm-level data from thousands of companies all suggest that there is a significant payoff from IT investments, contradicting the information paradox. However, these payoffs are contingent on a firm's ability to effectively adapt through organizational change to the new technology, and on a firm's ability to effectively manage the overall portfolio of IT investments. These results suggest that investing in IT is on average a positive ROI activity, but the benefits of IT investments are difficult to measure and risk factors can significantly impact the actual ROI realized.

REVIEW OF BASIC FINANCE

In this section we review the basic finance necessary to calculate ROI. The key concepts are the time value of money and internal rate of return (IRR). For a complete introduction to corporate finance see Brealey and Myers (1996). In the following section, a general framework is given for ROI analysis, and the ROI is calculated for a case example e-business project. The reader should note that ROI analysis for e-business investments and IT is in principle no different from ROI analysis for other firm investments such as plant and equipment, research and development, and marketing projects. All use the same financial tools and metrics and follow the general framework discussed in the next section.

The Time Value of Money

As an example, consider two e-business investments. Assume that both projects cost the same, but the first (Project 1) will have new revenue or cost-saving benefits of \$5 million (M) each year for the next five years, and the second (Project 2) will have benefits of \$11 M at the end of the first and second years, and nothing after that. If we only have enough capital to fund one project, which of these e-business projects is worth the most cash benefit today?

We might argue that the first investment's cash flows are worth \$5 M times five years, which is \$25 M, and the second project's payouts are \$11 M times two years, or \$22 M. From a purely financial perspective, assuming all other factors are equal, we would conclude by this reasoning that we should invest in the first project instead of the second. However, intuitively we know that \$1 today is worth more than \$1 in the future—this is the "time value of money." The dollar today is worth more because it can be invested immediately to start earning interest. So just adding the cash flows ignores the fact that \$5 M received today has more value than \$5 M received five years from now.

The correct approach is to discount the cash flows. That is, \$1 received in one year is actually worth $\$1/(1+r)$ today, where r is called the discount rate and is the annual interest rate investors demand for receiving a later payment. In this example, if r is 10%, a dollar received in one year is worth $\$1/1.1 = 91$ cents today. Similarly, cash received two years from now should be discounted by $(1+r)^2$, so that the dollar received two years in the future is worth $\$1/(1.1)^2 = 83$ cents today.

This argument can be generalized to a series of cash flows $A_1, A_2, A_3, \ldots, A_n$ received in time periods 1, 2, 3, ..., n. The value of these cash flows today is calculated from the discounted sum

$$PV = A_1/(1+r) + A_2/(1+r)^2 + A_3/(1+r)^3 + \cdots + A_n/(1+r)^n. \quad (1)$$

where n is the number of time periods and PV is called the present value of the cash flows. Discounting a series of cash flows is mathematically equivalent to weighting cash received in the near term more than cash received further in the future. The effect of inflation is generally ignored in the cash flows, so that $A_1, A_2, A_3 \ldots, A_n$ are given in today's prices. Inflation can be included in the present value calculation by adding an inflation factor to the discount rate. This is particularly important in economies that have high inflation rates. For a complete discussion of how to incorporate inflation see Brealey and Myers (1996).

In general, the series in Equation (1) can easily be calculated using the built-in present value function in personal computer spreadsheet software (such as Microsoft Excel) or using a financial calculator. For the special case when the cash flow is the same for each period ($A_n = A$), such as in a bank loan, the sum can be calculated in closed form:

$$PV = \sum_{k=1}^{n} \frac{A}{(1+r)^k} = A\left[\frac{1}{r} - \frac{1}{r(1+r)^n}\right]. \quad (2)$$

Returning to our original example, the present value of the two cash flows is calculated in Figure 3a assuming $r = 10\%$. In this example, PV(Project 1) = \$19 M and PV(Project 2) = \$19.1 M, so the expected cash benefits of the second project actually have more value today in present value terms than the first project. If the projects cost the same to execute, and this cost is less than \$19 M, a manager should prefer to invest in Project 2.

In order to compare projects that have different costs (investment amounts), it is useful to subtract the initial investment costs I from the present value, thus obtaining the net present value (NPV):

$$NPV = PV - I. \quad (3)$$

If the costs of the project are spread out over multiple time periods, then I is the present value of these costs. Hence from Equation (1), Equation (3) is equivalent to

$$NPV = -C_0 + \frac{(A_1 - C_1)}{(1+r)} + \frac{(A_2 - C_2)}{(1+r)^2} + \frac{(A_3 - C_3)}{(1+r)^3} + \cdots + \frac{(A_n - C_n)}{(1+r)^n}, \quad (4)$$

where the costs of the project $C_0, C_1, C_2, C_3, \ldots, C_n$ have been subtracted from the cash benefits $A_1, A_2, A_3, \ldots, A_n$ in the corresponding time periods 1, 2, 3, ..., n.

When making investment decisions, one always strives to invest in positive NPV projects. If the NPV of a project is negative, this means that the initial investment is greater than the present value of the expected cash flows. Investments in projects with negative $NPVs$ should not be

(a)

Project 1

	Year 0	Year 1	Year 2	Year 3	Year 4	Year 5
Final Payout Cash Flows		5	5	5	5	5
Present Value (US $ million)	**19.0**					

Project 2

	Year 0	Year 1	Year 2	Year 3	Year 4	Year 5
Final Payout Cash Flows		11	11	0	0	0
Present Value (US $ million)	**19.1**					

(b)

Project 1

	Year 0	Year 1	Year 2	Year 3	Year 4	Year 5
Final Payout Cash Flows		5	5	5	5	5
Initial Investment	(9)					
Present Value (US $ million)	19.0					
Net Present Value (US $ million)	**10.0**					
Profitability Index	1.11					

Project 2

	Year 0	Year 1	Year 2	Year 3	Year 4	Year 5
Final Payout Cash Flows		11	11	0	0	0
Initial Investment	(10)					
Present Value (US $ million)	19.1					
Net Present Value (US $ million)	**9.1**					
Profitability Index	0.91					

Figure 3: (a) The present value (PV) of Project 1 and Project 2 cash flows. (b) The net present value (NPV) and profitability index calculation. The discount rate is 10% for both (a) and (b).

made, because they do not add value to the firm and actually extract value.

Returning to our example, assume that the initial cost of Project 1 is \$9 M and the initial cost of Project 2 is \$10 M. From Figure 3b the *NPV*(Project 1) = \$10 M and *NPV*(Project 2) = \$9.1 M. Hence both projects have positive *NPV*, and should add value to the firm. However, if capital is limited (or rationed) one must select investments that have the most "bang for the buck." In other words, one must select projects that have the greatest returns for a given dollar of investment. A useful ratio capturing this idea is called the profitability index:

$$\textit{Profitability Index} = \frac{\textit{Net Present Value}}{\textit{Investment}}. \qquad (5)$$

For our example in Figure 3b, the profitability indices are 1.11 and 0.91 for Project 1 and Project 2, respectively, and *NPV*(Project 1) = \$10 M > *NPV*(Project 2) = \$9.1 M. Because the profitability index is greater for Project 1 than Project 2, if the funding decision is based purely upon financial metrics Project 1 is the preferred investment.

The present value and net present value clearly depend upon the discount rate. What discount rate should we use for an e-business investment? The discount rate used for investments in a specific firm is defined by the expected return of the combined debt and equity of the firm for a given industry. This discount rate is called the weighted average cost of capital (WACC) of the firm. Calculating the WACC for a firm is beyond the scope of this chapter; the interested reader is referred to Brealey and Myers (1996). However, as a rule of thumb, discount rates typically range from 10% to 25%, and a WACC of 15% or more is common in the technology industry. The Chief Financial Officer's (CFO's) office in a large company will usually calculate the WACC for use in investment decisions.

The discount rate is related to the risk of an investment so that firms in high-risk industries (such as technology) have higher WACCs—these companies in turn have higher expected returns in the stock market. Due to this risk–return relationship, the discount rate for more risky technology project investments is sometimes increased relative to that for less risky investments when *NPV* is calculated. A potential issue with this approach is that the discount rates chosen for riskier projects can be somewhat arbitrary. Arbitrarily increasing the discount rate adds additional uncertainty into the *NPV* calculation and may reduce one's objectivity in comparing projects. A better approach for technology investment decision-making incorporating project risk, and other factors such as the business value of the project, is discussed in the Executive Insights section.

The CFO's office will often compare investments based upon *NPV*, because this makes possible objective comparison and selection of the most profitable investments. The CFO is most likely managing a large portfolio of investments, and the power of the *NPV* approach is that is takes the guesswork out of financial decision making by placing all investments on a common footing. One limitation of *NPV* is that it does not take into account management flexibility to defer decisions into the future. The value of this management flexibility, or option value, is discussed in the Executive Insights section.

ROI, Internal Rate of Return (IRR), and Payback Period

Return on investment was defined in the Introduction as

$$ROI = \frac{Project\ Outputs - Project\ Inputs}{Project\ Inputs} \times 100\%. \quad (6)$$

where the project outputs are all of the benefits of the project quantified in terms of cost savings and revenue generation, and the project inputs are all of the costs of the project. The major problem with this definition is that it does not include the time value of money.

Specifically, ROI, defined by Equation (6), is rather vague, because a 100% ROI realized one year from today is more valuable than a 100% ROI realized in five years. In addition, the costs of the project may vary over time, with ongoing maintenance and professional services support. The benefits of the project may also vary over time, so that the cash flows are different in each time period. Equation (6) is therefore not a convenient way to compare projects when the inputs and outputs vary with time, and it is also not useful for comparing projects that will run over different periods of time. Due to these deficiencies, one typically uses internal rate of return (*IRR*) (Brealey & Myers, 1996). For good management decisions the ROI defined rather loosely in Equation (6) should translate in practice into calculating the IRR of a project's cash flow.

What exactly is IRR? The IRR is the compounded annual rate of return the project is expected to generate and is related to the *NPV* of the project, defined in Equations (3) and (4). The IRR is the discount rate at which the *NPV* of the project is zero. That is, the IRR is the discount rate where the cash benefits and costs exactly cancel. From this definition, the internal rate of return is calculated by solving for IRR in

$$NPV = -C_0 + \frac{(A_1 - C_1)}{(1 + IRR)} + \frac{(A_2 - C_2)}{(1 + IRR)^2} + \frac{(A_3 - C_3)}{(1 + IRR)^3} + \cdots + \frac{(A_n - C_n)}{(1 + IRR)^n} = 0 \quad (7)$$

where $A_1, A_2, A_3, \ldots, A_n$ are the positive cash benefits and $C_0, C_1, C_2, C_3, \ldots, C_n$ are the costs of the project in each time period 0, 1, 2, 3,..., n. In practice one most often uses spreadsheet software, or a financial calculator, and the built in *IRR* and *NPV* functions for calculations.

How do we make financial management decisions using *IRR*? When the *IRR* is greater than the project discount rate, or WACC, we should consider accepting the project—this is equivalent to a positive *NPV* project. When the *IRR* is less than the WACC the project should be rejected, because investing in the project will reduce the value of the firm. The tenet of basic finance theory is that all projects that have positive *NPV*, or *IRR* > WACC, should be funded. This is based upon the assumption that the firm has unlimited capital and, because positive *NPV* projects have an *IRR* better than the WACC of the firm, accepting these projects will increase shareholder value. As discussed in the previous subsection, however, in practice capital is limited (or rationed) and managers must make decisions based upon limited resources. The profitability index, Equation (5), can be used to calculate which projects have the greatest return per investment dollar. Hence positive *NPV* (or good *IRR*) is only one factor to consider in a technology investment decision.

Another concept that is a useful tool when combined with *IRR* and *NPV* is that of payback period. The payback period, or payback, is the time it takes for the project to recoup the initial investment. The payback period is calculated by cumulatively summing the net cash flows (projected revenues and cost savings less costs) of a project. When the sign of the cumulative sum of the net cash flows changes from negative to positive the project has "paid back" the initial investment. (For an ROI analysis where a new project is compared to a base case, without the project, the payback should actually be calculated from the incremental cash flows. See the case example in the following section.)

The payback period for a typical e-business project can be in the range of six months to two years, depending upon the type of project. It is unusual for an e-business project to have a payback period longer than two years. In making investment decisions, projects that have good *IRR* and the shortest payback periods are most often selected.

This section on introductory finance did not include tax or depreciation in the *IRR* analysis. The reader should note that the financial metrics *PV*, *NPV*, and *IRR* calculated with and without tax and depreciation can be very different. Tax and depreciation are important factors and are incorporated into the case example discussed in the following section.

In summary, return on investment analysis for technology projects is the process of calculating the *IRR* for a project. The calculation of *IRR* is based upon sound financial theory and is related to the *NPV* of the project. *NPV* and *IRR* are equivalent ways of incorporating the time value of money into financial investment decisions. In the following section these concepts are applied to an example e-business project and a template is given that is applicable to any technology *IRR* calculation.

CALCULATING ROI FOR AN E-BUSINESS PROJECT

The overall process of calculating *IRR* for a new project business case is straightforward. The first step is to calculate the base case revenue and costs expected in the future if the business continues as it is now. The next step is to calculate the net cash flows with the new proposed project; this includes total revenue, potential cost savings, and all

costs of the project. Finally, the base case cash flows are subtracted from the projected cash flows with the new project. The results of these subtractions are called the incremental cash flows for the project. The *IRR* is then calculated from these incremental cash flows. An equivalent approach is to calculate the additional benefits of the project directly to obtain the incremental cash flows. For complex business models, however, separating out the additional benefits when there are multiple variables can be more difficult than calculating the total cash flows with the new project and then subtracting the base case.

As discussed in the previous section, if the *IRR* calculated from the incremental cash flows is greater than the project discount rate, or WACC, the project should be considered for funding—this is equivalent to a positive *NPV* project. The challenge is to accurately incorporate the business drivers in the base case and all of the project costs, potential cost savings, and potential revenue benefits in the new project's cash flows.

In order to put the ROI calculation process in context, and to discuss some of the important details, it is useful to walk through an example. This section discusses a case example of ROI analysis applied to a Web portal e-business project. The Web portal in this example is a Web site with a product catalog, and customers can buy products and transact orders using the portal. The Web portal front end acts as a customer interface and, for a large firm, is typically connected internally to the firm's back-end IT systems, such as an enterprise resource planning (ERP) system, and other enterprise systems, such as customer relationship management (CRM) software.

The particular example discussed in this section is for a midsize electronics manufacturing company with global sales and operations. The example has been simplified to illustrate the main features of ROI analysis, and all numbers have been changed for confidentiality reasons. The cost and revenue numbers in this example are therefore for illustrative purposes only. The objective of this case example is to illustrate the general process and the important mechanics for calculating ROI rather than the exact costs and benefits of a Web portal project. For a detailed discussion and analysis of ROI for a Web portal e-business initiative and for an example of management of a Web portal development project see the two case studies in the references (Jeffery, et al., 2002a; Jeffery, et al., 2002b).

Base Case

The first step in setting up any ROI analysis is to understand the base business case. That is, what are the primary costs and revenues expected if the firm continues operations and does not implement a new e-business solution? Answering this question should focus on the major costs and revenue drivers that the new technology project is expected to impact. The process of understanding the existing business is called business discovery.

A best practice of business discovery is to understand the cost and revenue drivers in a particular business process and then benchmark against competitors in the industry. For example, if the average transaction cost for order processing in a firm is $35 per order, and the industry average is $10 per order, there is clearly an opportunity for improvement. Similarly, if the industry average take-rate (fraction of customers who accept a marketing offer) is 3% and a firm has a take-rate of 1%, there is an opportunity for improvement.

If e-business or other information technology is used by competitors to achieve cost or revenue improvements, benchmarking data provide estimates of the improvements that might be expected if a similar solution were applied to existing processes within a firm. Benchmarking data for IT are provided by several consulting groups. Because consulting services are most often the source of benchmarking data, one must be cautious that these data are accurate and applicable.

Understanding the key business drivers, and which factors can improve business performance, is essential and can have important bottom-line implications. For example, a major U.S. general retailer with over $40 billion in revenues used a Teradata enterprise data warehouse (EDW) combined with analytic CRM software to improve the target marketing of 250,000 catalogs mailed to customers each year. This initiative resulted in 1% improvement in the number of trips to stores generated among mailed customers, 5% improvement in the average purchase dollars per trip, and 2% improvement in gross margin, as the products featured in the advertisements for specific customer segments captured sales without reliance on "off-price" promotions. The initiative ultimately resulted in an increase in mailer revenue of $215 M per year, and the catalog targeting project alone with the new EDW and CRM technology had an *NPV* exceeding $40 M.

For the case example discussed in this chapter we can assume that the business discovery yielded a set of assumptions that are summarized in Figure 4. Specifically, the revenue and cost drivers are assumed to be the sales transactions to 1,700 customers and the transaction costs for processing these orders, respectively. The average sales revenue per order is $258, the average cost of goods sold (COGS) is 70% of each order, and the transaction cost

General Assumptions	
Discount rate (WACC):	12%
Tax rate:	35%
Customers in Year 0:	1,700
Transactions in Year 1:	141,000
Average order size in Year 1:	$258
COGS as a % of the sales price:	70%
Average order size annual growth rate:	3%
Base Case	
Number of transactions annual growth rate:	3%
Average processing cost per order:	$30
With the Web Portal	
Initial implementation cost:	$5M
Ongoing maintenance and marketing each year:	$1M
Jump in total transactions in Year 1:	20,000
Number of transactions annual growth rate after Year 1:	10%
Average processing cost of a Web transaction:	$3
Average processing cost per order:	$16.50
% total transactions with the Web portal in Year 1:	50%

Figure 4: Assumptions for the Web-portal case example.

using phone and fax averages $30 per transaction. In the next year (Year 1) the company anticipates 141,000 total transactions through existing channels and without a Web portal. Multiplying the average revenue per order by the number of transactions, and subtracting COGS and transaction cost, one can calculate the net income in Year 1. If the tax rate is 35%, the net Year 1 after-tax free cash flow is expected to be $4.3 M.

Cash flows projected into additional future years can be estimated by multiplying the Year 1 numbers by anticipated annual growth rate factors. One must make assumptions based upon the expected increase in sales and costs for the next few years. As part of the business discovery, these assumptions may be based on data for the firm's performance in the past. For simplicity in the present example we can assume that the firm is in a mature industry and anticipates 3% growth in the total number of transactions, assuming the Web portal imitative is not implemented. The base case three-year future (also called pro forma) cash flows derived from these assumptions are given in Figure 5a.

Note that this base case is simplified for this example and in practice may be much more complicated. For example, the revenue may come from multiple market segments with different transaction costs, and the number of transactions may be very large. See the references (Jeffery et al., 2002a ; Sweeney, et al., 2002a; Sweeney et al., 2002b) for examples of market segmentation and business discovery for complex ROI analysis.

Incorporating the E-business Project

The Web portal case example has two primary business objectives: (1) enable self-service order entry by customers, thus reducing costs, and (2) enable access into a broader market for customers, potentially increasing revenues. In addition to these business goals, the Web portal has strategic value, because in the electronic components manufacturing industry a Web portal is becoming a requirement for conducting business.

The costs of a project are often the easiest component of the IRR analysis to quantify. These costs may include items such as hardware, software, license fees, programmers' time, professional services (consulting), project management, hosting fees, outsourced contractors, and ongoing operating expenses. IT managers strive to keep the total cost of ownership of new products and systems at a minimum.

Minimizing total cost of ownership is related to the build vs. buy decision for a new IT or e-business project. This is because custom-built applications can have high total cost of ownership over their useful life. A useful rule of thumb is that if less than 10% custom modification to a packaged enterprise application is necessary then it is generally cheaper to buy than build. Greater than 10% custom modification puts the cost of building vs. buying about even, because new version releases of the packaged software will require continual custom modifications.

Web portal technology was novel in the mid-1990s, but by 2001 several vendors were offering stable solutions. Hence, for this case example the best approach is most likely to integrate commercial off-the-shelf packaged applications with the firm's existing enterprise software systems. The major costs will most likely be integration with existing systems and infrastructure to support high availability (24/7 operation with little or no down time) across multiple geographic markets. The cost of outsourcing the system, versus keeping it in-house, may also be considered. Detailed costing and a work breakdown structure would be completed for the final project plan. Cost estimates can also be obtained from similar projects that have been completed in the past.

For the purpose of this example we assume the project cost is $5 M, with ongoing costs of $1 M in each year. The ongoing costs include maintenance, upgrades, license fees, and professional services. To help facilitate the second business goal the Web portal initiative must include a marketing campaign in target markets. For simplicity in this example, these marketing costs are assumed to be included in the ongoing costs of the project. In practice the marketing plan would contain detailed costing and would most likely be broken out into a separate line item in the cash flow statement.

The primary anticipated benefits, or outputs, of the Web-portal initiative are reduced transaction costs and increased revenue generation. The cost savings occur because phone and fax orders for this company average $30 per order, and electronic processing is anticipated to cost $3 per order. The revenue generation benefit is expected to come from the Web portal's ability to have a global reach, so that with targeted marketing more customers can access the firm's products without increasing the size of the sales force. Other benefits of this initiative include fewer errors in processing transactions, reduced time to process orders, improved information on customers, and improved customer satisfaction, because customers can place orders 24/7 and have access to up-to-date product data.

Accurately quantifying all of the benefits of an e-business or IT system is the most challenging part of any ROI analysis. In practice one can often quantify the major hard cost savings. Revenue growth is more difficult to estimate and must come from market research, industry data, and past experience. It is often not possible to quantify soft benefits such as customer satisfaction and strategic advantage. The analysis therefore typically includes cost savings and revenue generation that can be estimated, and unquantifiable soft benefits are not included. This means that the ROI calculated will potentially be less than the realized ROI including soft benefits. One must then subjectively consider the project's soft benefits and how important they are to the firm. An ROI analysis is only as good as the assumptions that go into the analysis. The best practices for incorporating assumptions into an ROI model are discussed in the following section.

The details of the financial analysis calculation including the Web portal are described as follows. See Figure 4 for the assumptions and Figure 5b for the complete cash flow statement. Please note that what is most important in this chapter is the structure of the overall analysis, not the specific details.

For the case example, the average transaction cost is the easiest benefit to quantify and is straightforward to calculate. For all of the transactions processed, 50% of the

(a)

Base Case (No Web Portal)

	Year 0	Year 1	Year 2	Year 3
Customers	1,700	1,751	1,804	1,858
Number of Transactions		141,000	145,230	152,492
Average Order Size (US $)		258	265	273
Baseline Revenue (US $ thousands)		36,308	38,519	41,658
COGS (US $ thousands)		25,415	26,963	29,161
Order Processing Cost		4,230	4,357	4,575
Net Income		6,662	7,199	7,923
Free Cash after Tax (US $ Thousands)		4,330	4,679	5,150

(b)

New Web Portal Initiative

	Year 0	Year 1	Year 2	Year 3
Customers	1,700	2,081	2,299	2,454
Number of Transactions		161,000	177,100	194,810
Average Order Size (US $)		258	265	273
Revenue (US $ thousands)		41,458	46,971	53,219
COGS (US $ thousands)		29,020	32,880	37,253
Total Order Processing Cost		2,657	2,922	3,214
Gross Profit		9,781	11,169	12,751
Costs of the Web Portal Initiative				
Upfront Costs	(5,000)			
Ongoing Maintenance/Marketing		(1,000)	(1,000)	(1,000)
Depreciation Expense		(1,667)	(1,667)	(1,667)
Net Income		7,114	8,503	10,085
Net Income (After tax)		4,624	5,527	6,555
Add back the depreciation		1,667	1,667	1,667
Free Cash (US $ Thousands)	(5,000)	6,291	7,193	8,222

(c)

Incremental Cash Flows

	Year 0	Year 1	Year 2	Year 3
Net Incremental Cash Flows	(5,000)	1,960	2,514	3,072

Net Present Value (US $ thousands)	941
Discount Rate	12%
Tax Rate	35%
3 yr Internal Rate of Return (IRR)	**21.9%**

(d)

Payback Period Calculation

	Year 0	Year 1	Year 2	Year 3
Net Incremental Cash Flows	(5,000)	1,960	2,514	3,072
Cumulative Cash Flows		(3,040)	(525)	2,546
Payback is in 3rd month of Year 3 ===>				0.17

Figure 5: Case example of ROI analysis: (a) The base case free cash, (b) the free cash calculated including the Web-portal initiative, (c) the incremental cash flows, IRR, and NPV calculation, and (d) the payback period calculation.

customers are assumed to use the Web portal and 50% are assumed to use fax and phone methods of ordering. The average total transaction cost is the weighted average of the number of transactions expected using the new Web-portal system (assumed to be 50% of total transactions) multiplied by the transaction cost of $3 for each electronic transaction and $30 for each phone and fax order: $0.5 \times (\$3 + \$30) = \$16.50$ per order. With a larger fraction of customers using the e-business system, the average transaction cost per order decreases significantly from $30.

For this case example, we assume that with the new portal, market penetration will increase and that there will be an initial jump in the number of total transactions

in Year 1 as the global customer base is enabled to do online transactions. With a 14% increase in transactions in Year 1 and a 10% yearly growth in the total number of transactions driven by the marketing campaign in Years 2 and 3, the effective growth in gross revenues is 13.3% per year. Because it costs only $3 to process an order using the Internet, in addition to revenue growth there is also a substantial cost savings of $2 M due to the reduced average transaction cost to process an order.

Figure 5b incorporates the revenue and cost savings of the new Web portal initiative into a pro forma cash flow statement. The upfront and ongoing costs of the new initiative are also included. The revenue generation is incorporated in the increased number of transactions, and the cost savings are encapsulated in the total order processing cost line of the cash flow statement Figure 5b. For the calculation of net income we subtract out the depreciation of the project, assuming a three-year straight line schedule.

In the United States, for tax reasons new IT projects cannot be expensed in the year they are capitalized. The hardware, software, and professional service costs must be depreciated using a five-year MACRS (modified accelerated cost recovery schedule). This is an accelerated depreciation schedule described in Stickney and Weil (2000). Although the accounting books may use MACRS, depreciation for ROI analysis is most often incorporated using three- or five-year straight line depreciation. Straight line is a conservative compromise, because it weights the expense equally in each year, whereas accelerated depreciation weights the capital expense more in the first few years than in the last. Once the system is operational, ongoing costs such as maintenance and professional service support can be expensed when they occur.

Off balance sheet and lease financing options are usually not incorporated into the cash flow statements for the ROI analysis with a new project. For capital budgeting, the base case and the case with the new project should be objectively compared, independent of how the project is financed. Leasing and off balance sheet financing can artificially improve the ROI, because the cost of the project is spread over time by the lease payments. A more conservative estimate is to assume the costs of the project are incurred up front, or at the same time as the costs are anticipated to actually occur. Once the project is accepted for funding the best method of financing should be chosen.

To calculate the free cash flow with the new project, the last step is to add back the depreciation expense to the net income after tax. The depreciation expense was included in the calculation of net income in order to correctly include the tax advantage of this expense. However, for the final free cash flows the total depreciation is added back to the net income, because depreciation is not a "real" expense that actually impacts the cash flows, other than for tax reasons.

Incremental Cash Flows and IRR

Once the pro forma base case and new-project free cash flows have been calculated, the calculation of IRR is straightforward. The base case cash flows are subtracted from the cash flows with the new Web project; these are the incremental cash flows. See Figure 5c. The incremental cash flows are the net positive or negative cash in each time period that occurs in addition to the base case. The *IRR* is calculated from these incremental cash flows.

Using spreadsheet software, the *NPV* and *IRR* of the project are calculated by applying Equations (3) and (7), respectively, to the incremental cash flows. For the parameters given in this example, the NPV is $941,000 and the IRR is 22%, with a $5 M initial investment. Assuming the assumptions are correct, the IRR being greater than the firm's discount rate (WACC) suggests that this is a project the firm should consider funding.

Another factor to consider is the payback period. The payback for this project is calculated in Figure 5c from the incremental cash flows and occurs early in the third year (the beginning of the third month). The payback is anticipated to be just over two-years, which is potentially a little long, so one possibility is to consider adjusting the total project expenses to enable earlier payback.

The reader should note that if the major project expenses occur up front, and the net cash flows in later time periods are increasing and positive, the IRR will increase if the time period of the analysis is extended. For this case example, if the assumptions were extended into Years 4 and 5, the five-year IRR would be 46%, compared to 22% IRR for three years. This is because we have extended the time over which the cash benefits can be included in the calculation from three years to five, for the same up-front implementation cost.

Because the Web portal projects may produce benefits over a long time period into the future, an important question is, "What time period should be taken for a particular IRR calculation?" The time period for the analysis should match the time period used to calculate IRRs for similar investments in the firm. Often the one-, two-, and three-year IRR numbers are calculated for an investment decision, and depending upon the firm, management decides which one to use for comparisons with other projects. For the Web portal project example, 36 months was chosen as the length of time for the analysis. For e-business projects IRRs for time periods longer than three years are usually not considered when projects are compared, even though the project may have benefits in additional years.

Note that the 22% *IRR* calculated in this example does not include additional benefits such as: fewer errors in processing transactions, reduced time to process orders, improved information on customers, and improved customer satisfaction because customers can place orders 24/7 and have access to up-to-date product data. One can attempt to quantify these benefits and include them in the model; however, soft benefits such as improved customer satisfaction and better information are extremely difficult to accurately quantify. The approach most often used is to realize that the calculated *IRR* does not include these benefits, and hence the actual *IRR* of the project should be somewhat higher.

In addition, the case example does not include the strategic value of the initiative. Specifically, the Web portal may be a "table stake"—an investment that is required to stay in business in a particular industry. Hence, even if the *IRR* is less than the hurdle rate for the company, management must invest in the project, or risk losing market share to competitors who have the technology.

The complete ROI analysis for the case example e-business project is summarized in Figures 5a–5d. This spreadsheet can be used as a basic template and starting point for any technology ROI calculation.

UNCERTAINTY, RISK, AND ROI

As with any ROI analysis, the three-year IRR calculated at 22% in Figure 5c is only as good as the assumptions that are the foundation for the model. In this section we discuss how the assumptions and potential risk impacts of the project are essential factors to examine so that the ROI analysis supports the best possible management decision. The major uncertainties will come from the business assumptions and the risks of the technology project. We first focus on major uncertainties, business risks, and sensitivity analysis, and then on specific risks related to the technology. How to interpret ROI results and incorporate uncertainty and risk into the ROI analysis is also discussed.

Uncertainty

For the case example described in this chapter we know one thing for sure: the 22% *IRR* calculated in Figure 5c will not be the actual *IRR* obtained by the project. How do we know this? There are many assumptions that went into the simple analytic model, and there are risks that may impact the project. It is therefore practically impossible that the assumptions will indeed be exactly correct. The important realization is that the ROI analysis of Figure 5 is only a point estimate. Management decisions based upon this single estimate will not be as informed as decisions based upon a range of possible outcomes.

In creating the ROI analysis, there are several important questions to ask, such as: What are the major assumptions in the model? Does the model capture the essential drivers uncovered in the business discovery? What are the ranges of possible outcomes for each major assumption?

For complex problems, a simple yet effective method is to estimate the best, the worst, and the most likely case for each of the major assumptions. Market research, the business discovery, industry experience, and project management experience should be used to define a reasonable range of possible outcomes. The expected value of the *IRR* can then be estimated from (Project Management Book of Knowledge [PMBOK], 2003)

$$\text{Expected Value} = \frac{\text{Best Case} + 4 \times \text{Most Likely Case} + \text{Worst Case}}{6}. \quad (8)$$

Equation (8) is equivalent to weighting the best and worst cases individually by the probability .167 and the expected case by the probability .67 (the probabilities for approximately plus or minus one standard deviation for a normal distribution). If similar projects have been undertaken in the past, it may be possible to assign empirical probabilities to the best, worst, and most likely cases.

The best and worst case ROI numbers are just as important for the management decision as the expected value. The expected value is a point estimate of the most likely outcome, and the worst case *IRR* is an indicator of the downside risk of the project. Even with a good potential upside, funding a project that has a large downside risk of a very low or negative ROI can be questionable. If there is a wide variation of the best and worst case IRRs from the expected value, this is an indicator that there is significant risk in the project.

Equation (8) is a simple estimating tool to define the expected value of the ROI given a range of possible outcomes and is used in project management (PMBOK, 2003) to estimate the expected value of the cost and time for an IT project. Spreadsheet software enables sensitivity analysis of ROI models. This is a powerful and more sophisticated tool to help understand which parameters in a model are most important, and how these parameters interact.

Sensitivity Analysis

For the case example, the major assumptions in Figure 5 are the following:

The increased transactions as a result of the Web portal and the marketing campaign.

The fraction of existing customers who will migrate to use the Web portal over time.

The reduced transaction cost with the Web portal.

The cost of the project.

Two of these assumptions are particularly aggressive. First, we assume that when the Web portal becomes active 50% of the existing customer base will use the portal for transactions in the first year. The large number of users migrating to the system is the driver for the large cost savings. In practice the 50% migration may take longer than one year.

The second major assumption is that the number of transactions will jump by 20,000 in the first year, as a result of the global reach of the new Web portal, and that these transactions will then grow at a rate of 10% per year. This new revenue will not be possible without a significant and coordinated marketing campaign. Hence, this revenue generation assumption must be benchmarked against market research data and the experience of the marketing team.

Spreadsheet software (such as Microsoft Excel) enables one to dynamically change one or two variables in a model simultaneously and calculate the corresponding *IRR*. This analysis is surprisingly easy to do and provides a visual picture of the dependencies in any model. Figure 6a is the table of *IRR* output calculated by varying the total cost savings and the revenue generation. The "Auto Formatting" function enables color-coding of cells—gray was chosen for *IRR*s less than the hurdle rate of 12%, white for *IRR* greater than 12%. The gray cells correspond to cost saving and revenue generation amounts that would not be acceptable (negative *NPV*). The boundary, where the cells change from gray to white, is the minimum cost saving and revenue generation necessary so that the *IRR* approximately equals the hurdle rate ($NPV = 0$). These tables can be used as a tool to review the ranges of *IRR* in the context of the best, worst, and average cases expected for each input parameter.

(a)

Revenues (US $ thousands) \ Cost Savings (US $ thousands)	1,700	1,800	2,000	2,200	2,400	2,600	2,800
39,250	-26.3%	-25.8%	-24.7%	-23.7%	-22.6%	-21.5%	-20.3%
39,500	-20.3%	-19.7%	-18.7%	-17.6%	-16.4%	-15.3%	-14.1%
39,750	-14.6%	-14.0%	-12.9%	-11.8%	-10.7%	-9.5%	-8.3%
40,000	-9.2%	-8.6%	-7.5%	-6.3%	-5.2%	-4.0%	-2.7%
40,250	-4.0%	-3.4%	-2.3%	-1.1%	0.1%	1.3%	2.6%
40,500	1.0%	1.6%	2.8%	4.0%	5.2%	6.4%	7.7%
40,750	5.8%	6.4%	7.6%	8.9%	10.1%	11.4%	12.7%
41,000	10.5%	11.1%	12.4%	13.6%	14.9%	16.2%	17.5%
41,250	15.1%	15.7%	17.0%	18.3%	19.5%	20.8%	22.2%
41,500	19.6%	20.2%	21.5%	22.8%	24.1%	25.4%	26.8%
41,750	24.0%	24.6%	25.9%	27.2%	28.6%	29.9%	31.2%
42,000	28.3%	29.0%	30.3%	31.6%	32.9%	34.3%	35.7%
42,250	32.6%	33.2%	34.5%	35.9%	37.2%	38.6%	40.0%
42,500	36.8%	37.4%	38.8%	40.1%	41.5%	42.9%	44.3%

(b)

% of total customers migrating to the new Internet channel \ Lift in Transactions due to the Web Portal Initiative	4,000	6,750	9,500	12,250	15,000	17,750	20,500
25%	-17.0%	-13.5%	-10.1%	-6.9%	-3.7%	-0.6%	2.4%
29%	-13.5%	-10.1%	-6.7%	-3.5%	-0.3%	2.8%	5.8%
33%	-10.1%	-6.7%	-3.4%	-0.2%	3.0%	6.0%	9.1%
37%	-6.8%	-3.5%	-0.2%	3.0%	6.2%	9.3%	12.3%
41%	-3.6%	-0.3%	3.0%	6.2%	9.3%	12.4%	15.4%
45%	-0.5%	2.9%	6.1%	9.3%	12.4%	15.5%	18.6%
49%	2.6%	5.9%	9.2%	12.4%	15.5%	18.6%	21.6%
53%	5.6%	8.9%	12.2%	15.4%	18.5%	21.6%	24.7%
57%	8.6%	11.9%	15.1%	18.4%	21.5%	24.6%	27.7%
61%	11.5%	14.8%	18.1%	21.3%	24.4%	27.6%	30.6%
65%	14.4%	17.7%	21.0%	24.2%	27.3%	30.5%	33.6%
69%	17.2%	20.5%	23.8%	27.0%	30.2%	33.4%	36.5%
73%	20.0%	23.3%	26.6%	29.9%	33.1%	36.2%	39.4%
77%	22.7%	26.1%	29.4%	32.7%	35.9%	39.1%	42.2%

Figure 6: Case example of sensitivity analysis of the ROI model: (a) Cost savings versus revenues, and (b) percentage of customers shifting to the new Internet channel versus Year 1 transaction lift due to the Web portal initiative. Gray cells have IRR less than the 12% hurdle rate for the firm.

Figure 6b calculates the *IRR* as a function of two key drivers in the model: the number of new transactions and the fraction of customers using the new Web-portal channel. The boundary clearly shows the importance of migrating customers to the new channel to reduce transaction costs. Sensitivity analysis using the built-in functions in spreadsheet software (such as the Table function in Microsoft Excel) is a powerful tool to analyze the dependencies between variables in any ROI model.

Project and Technology Risks

A theme for this chapter is that the business drivers, rather than the specific technology, are often most important for any ROI analysis. However, risks of a technology implementation project can also have a significant impact on ROI. As discussed in the section on the productivity paradox, the majority of large IT projects fail to deliver on time and on budget (see Davenport, 1998; Rigby et al., 2002). The technology implementation project enters into the ROI analysis through the cost of the project and delays in realizing the revenue benefits, so that risk events often increase the cost and time of the project, decreasing the overall ROI. Risks for Internet projects and strategies to mitigate these risks are discussed in another chapter. Here we focus on specific risks that may impact the overall ROI of an e-business or IT project.

Keil and co-workers (Keil, Cule, Lyytinen, & Schmidt, 1998) conducted a research study of three panels of expert technology project managers in Finland, Hong Kong, and the United States. The three panels listed the common risk factors for any technology project in order of importance; see Figure 7.

What is so surprising about the list in Figure 7 is that managers across continents and in very different cultures perceive the same major project risks in order of importance. It is also interesting to note that technology is mentioned only once in this list—"Introduction of new technology" is third from the bottom.

1. Lack of top management commitment to the project
2. Failure to gain user commitment
3. Misunderstanding the requirements
4. Lack of adequate user involvement
5. Failure to manage user expectations
6. Changing scope/objectives
7. Lack of required knowledge/skills in the project personnel
8. Lack of frozen requirements
9. Introduction of new technology
10. Insufficient/inappropriate staffing
11. Conflict between user departments

Figure 7: Risk factors identified by three independent panels of technology project managers listed in order of importance. Adapted from Keil et al. (1998).

In the early and mid-1990s Internet technology was new and many new Internet technology projects of that time period were "bleeding edge." These new Internet solutions were much more complex than previous IT systems. In addition, the Internet mania and infusion of vast amounts of venture capital pushed product development to "Internet time" in order to grab market share (Iansiti & MacCormack, 1999). These time pressures resulted in buggy code releases, and beta versions abounded. ROI for such new technology projects where costs and benefits were relatively unknown, was very difficult to define.

However, in 2003 and beyond, with Internet technology entering the mainstream and distributed architectures becoming more the norm than the exception, practically all technology investments are required to demonstrate a good ROI. Fairly good and systematic cost estimates for e-business systems are available today. The business benefits of these systems, although still difficult to quantify, are easier to estimate than when the technology was first introduced.

From Figure 7, the primary project risk factors are therefore not technological but organizational. For example, the top two risks in the list Figure 7 are "lack of top management commitment" and "failure to gain user commitment." These risk factors involve the people who will support and use the project and are risk factors that a project manger has little or no control over. Organizational issues are an essential consideration for the success of any technology project. Figure 7 is a simple tool one can use to assess the major risks of a project that may impact the ROI. If any of these risk factors are present, they should be included at least qualitatively in the management decision. In addition, a risk management strategy can be invaluable for planning contingencies for mitigating various risk events (Karolak, 1996).

Monte Carlo Analysis Applied to ROI

Sensitivity analysis using spreadsheet software is a useful tool for visualizing the interrelationships between parameters in an ROI model. However, this method has the limitation that one can vary at most two parameters simultaneously. Even for the relatively simple model given as a case example in this chapter, several parameters combine to give the ROI. The variation of multiple parameters simultaneously can be included using Monte Carlo methods.

The idea of a Monte Carlo simulation is to generate a set of random numbers for key variables in the model. The random numbers for a specific variable are defined by a statistical distribution. Similarly to defining the best, worst, and expected case for each input parameter in a sensitivity analysis, the shape of the distribution and spread (mean and standard deviation) are best defined by the management team. Past experience, market research, and the judgment of the management team are all factors to consider when defining the statistics of the input variables.

The random numbers are then put into the analysis spreadsheet and the output (the *IRR* and *NPV*) is calculated. A new set of random numbers is then generated based upon the statistical functions defined for each input variable, and the output is recalculated. If this process is repeated a large number of times statistics can be generated on the output of the model. Intuitively, one Monte Carlo cycle is a possible outcome of the model with one particular set of variations in the inputs. By running thousands of cycles, one is effectively averaging what might happen for thousands of identical projects given many different variations of input parameters.

Relatively low-cost packaged software is available that can perform Monte Carlo simulations in spreadsheet software (Crystal Ball 2003, Palisade @Risk 2003). This software is easy to use—the user selects specific cells and specifies distribution functions for the variables. The software then varies the values of the cells with random numbers. The output, in this case the IRR or NPV, is automatically calculated for a large number of cycles and statistics of the possible outcomes are generated.

Figure 8 is an example of the Monte Carlo output for the case example of Figure 5. The project cost, increase in number of transactions, and percentage of users migrating to the Web channel were varied simultaneously. The distribution functions chosen for the inputs were all normal distributions with standard deviations $1 M, 15,000, and 25%, respectively. The average IRR, or expected value, is 22%, with standard deviation 17.5%.

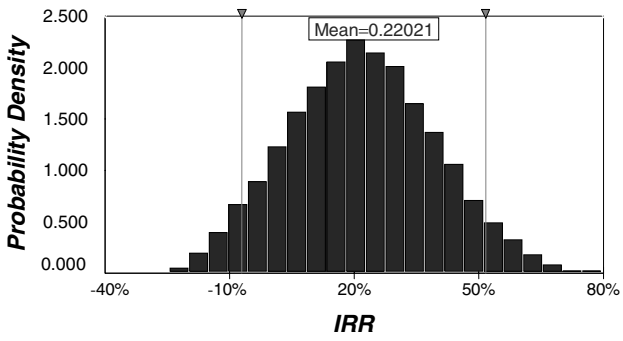


Figure 8: Distribution of three-year IRR calculated from 10,000 Monte Carlo iterations.

The Monte Carlo analysis shows that the model has considerable spread in the *IRR* with these parameters. Specifically, there is a 28% probability that the project will have an *IRR* less than the hurdle rate for the company. Given this information, the management team can consider whether they will fund the project as is, kill the project, or revise the scope and assumptions to reduce the downside risk.

EXECUTIVE INSIGHTS

This chapter has developed the tools necessary for calculating ROI for an e-business or IT project. This section provides a "big picture" framework for how ROI is used for technology investment decisions and what questions to ask when reviewing an ROI analysis. We also look "beyond ROI" at trends for the future.

The Important Questions to Ask When Reviewing an ROI Analysis

This chapter has discussed the major issues concerning ROI analysis and factors to consider in developing an analytic financial model for technology projects. The following set of questions summarizes the issues that were discussed. These questions may be useful to consider when reviewing an ROI analysis:

1. What are the main assumptions in the model?
2. Was there a business discovery to define the assumptions?
3. Are all the major uncertainties and risks adequately accounted for?
4. Are the assumptions realistic and are they expressed as a range of possible inputs?
5. Is the calculated IRR expressed as a range with an expected value and approximate probabilities?
6. Is there a sensitivity analysis and how is it interpreted?
7. What is the downside risk (worst case) and is there a plan to mitigate this risk?
8. Will the project have senior management and end user support, are the requirements well defined, and will an experienced project manager run the project?
9. What is the strategic value of the project to the firm in addition to the benefits incorporated in the model?
10. How important are other factors, such as soft benefits, that were not included in the analysis?
11. Does the project contain any option value that should be factored into the decision?

As described in detail in the section Risk, Uncertainty, and ROI, the analysis is only as good as the underlying assumptions. The first four questions are designed to probe if the assumptions incorporate the important issues, how they were obtained, and if the uncertainty in the assumptions is understood. Assumptions are critical to the validity of the ROI model. An effective method is for the management team to collectively define the assumptions based upon their experience and market research. If the assumptions are all based upon conservative estimates, and the management team collectively agrees on the assumptions, the ROI analysis is ultimately more convincing.

Questions 5 through 7 probe if the range of possible outcomes is understood and if there is a plan to deal with the worst case. Question 8 asks if the primary organizational risks have been thought through. In addition to Question 8 the list in Figure 7 can be used as a checklist for additional potential risks that may impact the project and Karolak (1996) gives a complete software project risk management checklist. Finally, questions 9 through 11 probe for additional value that may not have been captured in the ROI analysis and that should be considered for the funding decision.

The last question, 11, is concerned with the potential option value of the project—from the survey of Fortune 1000 CIOs 20% of respondents report that they qualitatively consider option value in funding IT projects (Jeffery & Leliveld, 2002). What is the option value of a technology project? An e-business or IT project has option value if, as a result of the project, the firm has the opportunity to implement additional projects in the future, and these projects would not have been possible without the initial project investment. Option value can be an important component of added value and is especially important for infrastructure investments.

For example, an enterprise data warehouse (EDW) is a very large IT infrastructure investment that, from a cost containment perspective, may be difficult to justify. However, once this infrastructure is put in place, the firm can leverage it for a variety of potential applications: Analytic CRM, improved supply chain management (SCM), and improved demand chain management (DCM) are a few of these applications. Hence, implementing the EDW is equivalent to buying options for CRM, improved SCM, improved DCM, and a variety of other strategic initiatives. Analytic methods exist for calculating financial option values and these methods have been applied to technology projects (McGrath & MacMillan, 2000). Qualitatively at least, the option value of a technology project should be considered when making an investment decision.

A Framework for Synchronizing E-business Investments With Corporate Strategy

A major challenge for executive managers is how to decide which new e-business and IT projects to fund. This is a complex decision, because for a large firm the annual IT budget may be several hundred million dollars or more and often there can be many new projects that must be considered for investment. For example in the 1990s a major worldwide banking institution, which was representative of other industry leaders, had an annual IT budget of $1.3 billion and had over 800 projects running simultaneously.

The process of managing the portfolio of technology investments of a firm is called IT portfolio management. This process is similar to managing other portfolios in the firm such as financial assets, new products, and marketing initiatives. IT portfolio management includes important factors such as the strategy of the firm and the risk and return of investments. This idea is not new and was first discussed by McFarlan (1981).

- Define the firm-wide strategic intent and business objectives
- Understand the strategic context of the firm. This context defines the focus of the technology investments
  - *Corporate strategy: operational excellence, customer focus, innovation*
  - *IT focus: Cost reduction, defined by strategy, strategy enabler*
- Develop e-business and IT objectives matched to the corporate strategic objectives
- Develop an appropriate portfolio of e-business and IT investments to support the strategic business objectives
  - *Make risk and return (ROI) tradeoffs on investments*
- Update as necessary
  - *Requires a continual dialogue of cross functional executives and technology managers*

Figure 9: Linking strategy to IT portfolio investments: a framework for managing IT by business objectives. Adapted from Weill and Broadbent (1998).

As discussed in the Introduction and throughout this chapter, ROI analysis is only one component of a technology investment decision. A general framework for investing in technology is given in Figure 9. This top-down approach (Weill & Broadbent, 1998; Weill, Mani, & Broadbent, 2002) starts with executive managers defining the strategic objectives of the firm. From the corporate strategy the key business objectives are defined. For example, these objectives may include increasing revenues in core markets, growing revenue in specific new markets, or cutting costs internally.

When defining the strategic initiatives, it is important to understand the strategic context of the firm within a given industry. The major focuses of corporate strategy can be grouped approximately into three categories: operational excellence, customer focus, or innovation. Treacy and Wiersema (1997) conducted a research study of thousands of firms and found that market-leading firms were often exceptional in one or two of these three categories, but none were exceptional in all three. One example is Dell Computer: Dell excels at operational excellence and customer service, but does not produce particularly innovative products. Another example is IDEO, a design company that has won countless awards for product innovation focused on what customers need.

In 2000 and beyond, the line between the three focuses of operational excellence, customer focus, and innovation is blurring. Increasingly, all firms must exhibit some level of customer focus excellence to remain competitive. However, understanding the core drivers of a firm's business is an essential first step to ensure that investment dollars are optimally allocated. The goal is to synchronize e-business and IT investments with the corporate strategy. The IT objectives for the firm must support the key business objectives (KBOs) derived from the corporate strategy in order to optimize the value of the portfolio of IT investments. Synchronization of IT with corporate strategy is simply not possible if the KBOs are not well defined.

Once the key IT objectives have been defined, the next step in the process in Figure 9 is to select an optimal portfolio of projects. This can be a challenging task, because often capital is limited and there may be many potential projects that could be funded. How do we select an optimal portfolio of e-business and IT investments? A rigorous IT portfolio management selection process can help capture the value of the project to the business and the risk of the project.

Kaplan and Norton (1992) have pioneered the use of scorecards to rate business performance. Scorecards are a powerful tool to objectively rank technology projects against one another. As an example, Figure 10 is the scorecard used by Kraft Foods to rank IT and e-business projects. Note that there are two dimensions of the scorecard: "Business Value Criteria" or value to the business, and "Likelihood of Success Criteria" or ability to succeed. Ability to succeed is related to the risk of the project. Also note that ROI, labeled as financial return, is just one component of the total score.

The categories on the scorecard and the category weights were defined by the Kraft Foods executive management team. A detailed grading rubric was developed

Likelihood of Success Criteria		Wt.	Score	**Business Value Criteria**		Wt.	Score
Technical Standards	X1:	10%		Financial Return	Y1:	30%	
Skills Capability & Training	X2:	10%		Customer & Consumer Focus	Y2:	20%	
Scope & Complexity	X3:	25%		Supply Chain Business Benefits	Y3:	15%	
Business Alignment	X4:	22%		Technology Efficiency	Y4	15%	
Risk Factors	X5:	21%		Knowledge Advantage	Y5:	10%	
Management Capability	X6:	12%		Work life Balance	Y6:	10%	
Dimension Total	X	100		Dimension Total	Y	100	

Figure 10: Kraft Foods score card used to rank new e-business and IT projects on the dimensions of ability to succeed and value to the business. Source: S. Finnerty, CIO of Kraft Foods and President of the Society for Information Management, (personal communication, December 2002).

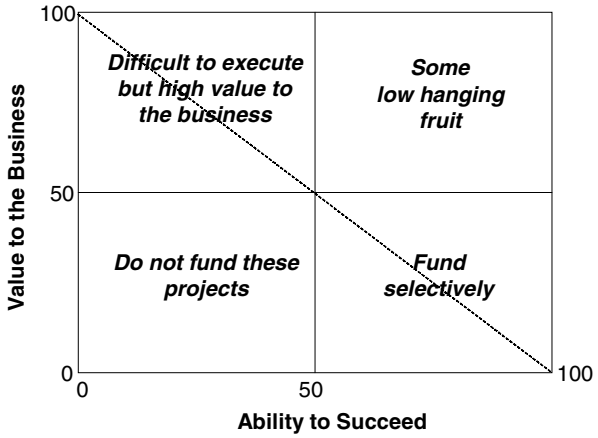


Figure 11: The portfolio application model.

so that each category could be objectively scored, and an independent review committee evaluated all projects and ensured consistency in scoring. All projects were then ranked by the business value criteria total score, and a line was drawn that corresponded to the total IT budget. The projects were also plotted on the portfolio application model matrix, Figure 11. The portfolio application model makes possible a schematic of the risk and return profiles for all of the IT projects. Based upon this information, the executive management team at Kraft Foods, which included the CFO and business unit sponsors, discussed which projects to fund and which to reject. The discussion enabled the CIO to increase the IT budget, with the CFO's approval, in order to fund additional projects that had high value to the business.

As a general example, if a KBO for a firm is to cut costs, a corresponding IT objective may be to increase electronic transaction processing. On the scorecard, projects that support electronic transactions will be weighted more than projects that do not. New e-business projects such as e-procurement are therefore more likely to be selected for funding through the IT portfolio management selection process. An e-procurement system may also be considered to have a relatively high ability to succeed, or equivalently a low risk.

Projects plotted on the matrix in Figure 11 fall into four categories. Projects in the upper right have high value to the business and ability to succeed. These projects should be funded. Small and medium-sized e-business projects such as e-procurement and customer self-service portals may fall into this category, and are often "low-hanging fruit," projects that will yield quick payback. Projects in the lower left corner have low value to the firm and have high risk—these projects clearly should not be funded.

Projects on the upper left in Figure 11 have high value to the company but are difficult to execute. Example projects may be ERP, CRM, or EDW and large strategic e-business initiatives. These projects may well be drivers for the long-term competitive advantage of the firm. Risk is clearly an issue with these projects, and a risk management plan can potentially significantly improve the ability-to-succeed score. In order to reduce the risk for a large project, the project may be broken into components or phases that each have a high ability to succeed. Projects that fall into the lower right corner in Figure 11 have low perceived value, but have a high ability to succeed. IT executives may choose to selectively fund projects in this category because they can be easy wins for the IT team.

A potential issue is that infrastructure investments may often be categorized as having low value to the business by non-IT business executives. The low value-to-the-business score may be due to the value not being accurately captured on the score card. Infrastructure is an important platform for future projects and may have significant option value. However, without a specific category for option value an infrastructure investment may receive a low value-to-the-business rating as perceived by executive managers. Future IT initiatives often depend on an infrastructure being in place. Therefore, for infrastructure projects the option value and future dependencies can be important considerations for the funding decision.

The IT portfolio management process gives executive managers a framework for optimal investment decision-making. Implementing this framework in practice gives managers objective information that can be used to make informed management decisions. Ultimately the management decision is made based upon executives' experience and must weigh subjective issues that are not quantified by the process. In addition, executives should also consider the dependencies between projects and the optimal order for execution. Kraft Foods exemplifies how a cross-functional executive team discussed the available information and reached consensus on the funding decision.

Finally, to effectively synchronize strategy and IT investments the IT portfolio management process must be ongoing. Many firms in mature industries have fixed annual IT budget cycles, so that the IT portfolio management process is implemented for the funding decisions of each cycle. However, in order to optimize the return from IT investment dollars, firms in dynamically evolving industries should implement quarterly or more frequent IT portfolio reviews.

Beyond ROI: Trends for the Future

Following the bursting of the Internet bubble in 2000, the technology industry is undergoing a shakeout and consolidation, which may last several years. As we look forward in this environment, optimizing investments in e-business and information technology is increasing in importance as companies struggle to maintain competitive advantage. Calculating ROI is important for informed management decisions. However, as we have discussed, ROI is only one component of the decision-making process.

The method of calculating ROI for an e-business or IT project is in principle no different from the method for calculating ROI for a new manufacturing plant, marketing plan, or research and development project. However, e-business and IT projects can be incredibly complex, so that estimates and generalities that are good enough for a manufacturing project can potentially destroy an IT project if any element goes wrong. Building the ROI model on sound assumptions and developing a risk management strategy can therefore significantly impact the actual ROI realized for IT projects.

A trend for the future will be that firms will increasingly implement more sophisticated IT portfolio management processes and will incorporate ROI into these processes. Furthermore, we have discussed ROI in the context of new project selection. In order to maximize IT value one must realize that ROI analysis is an important on-going process. That is, the ROI of projects should be measured after the project is complete. This after action review enables feedback to the entire IT portfolio management process, and the firm can then calculate the realized ROI of the entire IT portfolio.

Similarly to a financial portfolio, it does not make sense to invest in a mutual fund or stock that is losing money year after year. E-business and IT projects are no different, and measuring the ROI of existing IT projects enables executives to weed out underperforming investments.

Some complex strategic e-business initiatives may have high cost, high risk, and huge potential payoffs. For these projects a management strategy is to break the project down into phases, where each phases is defined by ROI. Once a phase is complete it should demonstrate good ROI before the next phase is funded. This approach reduces the risk of the e-business investment and makes the project "self-funding," because new revenue or cost savings can fund the next phase of the initiative.

During the roaring 1990s, Internet and e-business initiatives were viewed as too complex, or too innovative, for management investment decisions to be made using ROI. As we move into the next phase of the technology revolution powered by the microprocessor and networking technologies, e-business initiatives will be scrutinized and evaluated on the same basis as all other firm investments. IT management teams must therefore embrace the financial management techniques of ROI analysis and portfolio management that are used widely in other functional areas of the firm.

ACKNOWLEDGMENTS

The author gratefully acknowledges Sandeep Shah for his help preparing the manuscript and the ROI analysis with Monte Carlo simulations. He also thanks Professor Robert Sweeney at Wright State University and Joe Norton of the Norton Solutions Group for useful discussions.

GLOSSARY

COGS Cost of goods sold, equal to the beginning inventory plus the cost of goods purchased or manufactured minus the ending inventory. These costs are expensed because the firm sold the units.

DCF Discounted cash flow, equal to future cash flows divided by discount rate factors to obtain present value.

Depreciation The portion of an investment that can be deducted from taxable income. It is also the reduction in book market value of an asset.

Discount rate The rate used to calculate the present value of future cash flows.

Hurdle rate The minimum acceptable rate of return on a project.

Information technology portfolio management A methodology for managing information technology investments as a portfolio with different risks and returns. The process often involves using scorecards to rate projects on multiple dimensions, such as the alignment of the project with the strategic business objectives of the firm and the ability of the project to succeed.

IRR Internal rate of return, the discount rate at which the net present value of an investment is zero.

ITPM Information technology portfolio management.

MACRS Modified accelerated cost recovery system, the accepted U.S. income tax accelerated depreciation method since 1986.

NPV Net present value, a project's net contribution to wealth—present value minus initial investment.

Payback The payback period of an investment, or the time taken to recoup the original investment with the new revenue and/or cost savings from the project.

PV Present value, the discounted value of future cash flows.

Real option A deferred business decision that is irreversible once made and whose eventual outcome is contingent upon the future evolution of the business environment.

Risk free rate The expected return for making a safe investment, usually equivalent to the rate of return from government bonds.

ROI Return on investment, a generic term for the value of a project relative to the investment required. In practice the ROI for a project is calculated as the IRR for the project.

Table stake A technology investment that is necessary in order to remain competitive in a particular industry.

Time value of money The idea that cost savings or revenue received today is more valuable than the same cost savings or revenue received some time in the future.

WACC Weighted average cost of capital, the expected return on a portfolio of all the firm's securities. Used as the hurdle rate for capital investment.

CROSS REFERENCES

See *E-business ROI Simulations; Electronic Commerce and Electronic Business; Risk Management in Internet-Based Software Projects*.

REFERENCES

Brealey, R., & Myers, S. (1996). *Principles of corporate finance*. New York: McGraw-Hill.

Brynjolfsson, E. (1993). The productivity paradox of information technology. *Communications of the ACM, 36*(12), 67–77.

Brynjolfsson, E., & Hitt, L. (1996). Paradox lost? Firm-level evidence on the returns to information systems spending. *Management Science, 42*(4), 541–558.

Brynjolfsson, E., & Hitt, L. (1998). Beyond the productivity paradox. *Communications of the ACM, 41*(8), 49–55.

Chrystal Ball. Retrieved January 11, 2003, from http://www.decisioneering.com

Davenport, T. (1998, July–August). Putting the enterprise into the enterprise system. *Harvard Business Review,* 121–131.

Dehning, B., & Richardson, V. (2002). Returns on investment in information technology: A research synthesis. *Journal of Information Systems, 16*(1), 7–30.

Dewan, S., & Min, C. (1997). The substitution of IT for other factors of production: A firm-level analysis. *Management Science, 43*(12), 1660–1675.

Iansiti, M., & MacCormack, A. (1999). Living on Internet time: Product development at Netscape, Yahoo!, NetDynamics, and Microsoft. Harvard Business School Case 9-6967—52.

Jeffery, M., & Leliveld, I. (2002). Survey: IT portfolio management challenges and best practices. Retrieved January 11, 2003, from Kellogg School of Management and DiamondCluster International at http://www.kellogg.northwestern.edu/ITsurvey

Jeffery, M., Ritters, T., & Anfield, J. (2002a). B&K distributors: Calculating return on investment for a Web-based customer portal. Kellogg School of Management Case. Retrieved January 11, 2003, from http://www.kellogg.northwestern.edu/IT/portfoliocases

Jeffery, M., Yung, D., & Gershbeyn, A. (2002b). AD high tech case A: Managing projects for success. Kellogg School of Management Case. Retrieved January 11, 2003, from http://www.kellogg.northwestern.edu/IT/portfoliocases

Karolak, D. (1996). *Software engineering risk management*. Los Alamitos, CA: IEEE Computer Society Press.

Kaplan, R., & Norton, D. (1992). The balanced score card—Measures that drive performance. *Harvard Business Review, 70*(1), 71–79.

Keil, M., Cule, P., Lyytinen, K., & Schmidt, R. (1998). A framework for identifying software project risks. *Communications of the ACM, 41*(11), 76–83.

Malone, T. (1997, Winter). Is empowerment a fad? *Sloan Management Review, 38,* 2.

McFarlan, F. (1981). Portfolio approach to information systems. *Harvard Business Review, 59*(5), 142–150.

McGrath, R., & MacMillan, V. (2000). Assessing technology projects using real options reasoning. *Research–Technology Management, 43*(4), 35–49.

Palisade @Risk. Retrieved January 11, 2003, from http://www.palisade.com

Project management book of knowledge (2003). Retrieved January 29, 2003, from http://www.PMI.org

Rigby, D., Reichheld, F., & Schefter, P. (2002, February). Avoid the four perils of CRM. *Harvard Business Review,* 101–109.

Stickney, C., & Weil, R. (2000). *Financial accounting: An introduction to concepts, methods, and uses*. New York: Harcourt.

Strassmann, P. (1990.) *The business value of computers*. New Canaan, CT: Information Economics Press.

Sweeney, R., Davis, R., & Jeffery, M. (2002a). Teradata Data Mart consolidation return on investment at GST. Kellogg School of Management Case. Retrieved January 11, 2003, from http://www.kellogg.northwestern.edu/IT/portfoliocases

Sweeney, R., Davis, R., & Jeffery, M. (2002b). ROI for a customer relationship management initiative at GST. Kellogg School of Management Case. Retrieved January 11, 2003, from http://www.kellogg.northwestern.edu/IT/portfoliocases

Treacy, M., & Wiersema, F. (1997). *The discipline of market leaders*. New York: The Free Press.

Weill, P., & Broadbent, M. (1998). *Leveraging the new infrastructure—How market leaders capitalize on information technology*. Boston: Harvard Business School Press.

Weill, P., Subramani, M., & Broadbent, M. (2002, Fall). Building IT infrastructure for strategic agility. *Sloan Management Review,* 57–65.

Risk Management in Internet-Based Software Projects

Roy C. Schmidt, *Bradley University*

INTRODUCTION

Despite the success stories in the literature, it remains a sad statistic that too many software development projects end in failure. At least a quarter of all software projects are cancelled outright, and it is likely that 75% of all large systems projects are dysfunctional because the systems produced do not meet specifications or are simply not used. The average software project runs over its budget by 50% (Gibbs, 1994; Lyytinen & Hirschheim, 1987; van Genutchen, 1991; Walkerden & Jeffrey, 1997). It is common knowledge that software projects are risky, and thus there has been an intense search for appropriate managerial action to counteract software project risk.

One approach that has gained popularity in recent years is the use of the Internet as a vehicle for communication among project team members (Benett, 1998; Collaborative Strategies, 2000). By improving communications and facilitating work flow management, Internet-based project management seeks to reduce some elements of risk. A thriving industry of support and management of Internet-based projects has grown up in just five years. One such company is Reinvent Communications (formerly USWeb), which not only supports Internet collaboration but also uses its own tools to provide its support (Bumbo & Coleman, 2000). At this point, it is too early to tell if this new management approach has had any impact on project success, but the new approach may be displacing some risk factors with new ones, thus creating new opportunities for failure (Ash, 1998).

A more traditional method for reducing failure in software projects is the concept of software project risk management. Advocates of software project risk management claim that by identifying and analyzing threats to success, action can be taken to reduce the chance of failure. In this chapter, the traditional risk management approach is extended to include the new Internet-based project management environment.

Risk management is a two-stage process: assessing the risk and taking action to control risk. The first stage, risk assessment, consists of three steps: (a) identification of risk factors, (b) estimation of the likelihood for each risk factor to occur and potential damage from the risk, and (c) an evaluation of total risk exposure (Charette, 1989). So the first step toward control of software project risk is an understanding of the exact nature of risk factors.

WHAT IS RISK?

The concept of risk is not clear to most practicing managers. Much like the problems the average person has understanding Bayesian statistics, the word "risk" usually conjures up an image of negative outcomes. A "risky" project is seen as one that is likely to fail. A risk factor is seen as an opportunity to fail. This view of risk overlooks the true nature of the beast.

Risk as an Element of Decision Theory

The following explanation is highly simplified but sufficient for the purpose of this chapter. For a complete treatment of decision theory and risk, see the collection of essays by Arrow (1965).

Decision theory describes the decision-making process as a choice of actions leading to a specific outcome. A small number of actions lead to outcomes that are favorable, a few to those that are unfavorable, but the vast majority lead to outcomes that are unclear. Assuming that the clearly favorable outcomes are too small to be attractive, the unclear alternatives present the decision maker with risky opportunities.

For each of the risky alternatives, there is some probability of a favorable outcome. For example, a given alternative may have 15% chance of success (a favorable outcome) and thus an 85% chance of failure. How would this alternative compare with another that has 99% chance of success? To make this evaluation, one also needs to know the magnitude of the outcome.

Suppose that the first alternative, if successful, would net $1 million in profit but a failure would mean the loss of $50,000 invested in the alternative. The second alternative would net $500,000 profit, but failure would lead to a loss of $3 million. Although the second alternative seems unattractive because of the magnitude of the loss and the smaller profit, it is actually the better choice. To compare the alternatives, both the probability of the outcome and

its magnitude must be accounted for. The product of the magnitude and probability is called *expected outcome*. The first alternative promises a positive expected outcome of $150,000 ($1 million × 15%), or a negative expected outcome of $42,500 ($50,000 × 85%). The second alternative could lead to a positive expected outcome of $495,000 or a negative expected outcome of only $30,000. So despite appearances, the second alternative is less risky.

For this reason, risk is evaluated based on the product of the magnitude of an outcome and its probability of occurrence. A risk factor is any event that embodies risk. Risky alternatives have both positive and negative expected outcomes, and both must be considered when evaluating the risk. This approach to evaluating software project risk is strongly advocated by Charette (1989). Surprisingly, practicing managers often overlook these facts.

Risk as Viewed by Practicing Managers

There are two important biases affecting risk evaluation that have been observed among practicing managers. First, managers act as if they are evaluating only the magnitude of a loss, without considering the probability of the loss. Their focus is on the loss, not the potential gain, because risk carries a negative connotation to them. As a result, their ranking of alternatives is often out of line with expected outcomes (March & Shapira, 1987).

Second, managers do not attend to all potential risks. Regardless of magnitude or expected outcome, if a risk factor appears to be outside their direct control, managers assign them to a category that might be called "Acts of God." They make no attempt to ameliorate these risk factors, even if the expected outcome could be shutdown of a project (Schmidt, Keil, Lyytinen, & Cule, 2001).

Information Technology (IT) Project Risk

IT project risk factors have been defined as any factors that could affect the success of IT projects (Schmidt et al., 2001). Numerous studies have sought to catalog IT project risk factors. A thorough review of risk factors and a ranking of the most important risk factors are presented by Keil, Cule, Lyytinen, and Schmidt (1998) and Schmidt et al. (2001). All 53 of the risk factors discussed in these two papers apply to both traditional IT projects and Internet-based projects. There are significant differences between them, however. As a result, additional risk factors are associated with those differences. In the following section, some of those additional risk factors are examined.

INTERNET-BASED PROJECT RISKS

In recent years, the management of software projects, as well as other types of projects, has met with a new venue, the Internet. By taking advantage of the flexible, ubiquitous communications afforded by the Internet, project managers have been able to extend team membership to analysts and programmers who would not have worked on the project because of distance and time factors. Using groupware (both integrated commercial packages and stand-alone tools), Internet-based project teams can become larger, with more diverse interests and talents. Workflow management software allows project managers to track tasks and monitor team member contributions. Whether these new tools are better than traditional techniques is not well established. Even teams that share the same geographic location are attempting to benefit from the new Internet-based project management tools through the use of intranets (Benett, 1998).

How Internet-Based Projects Differ From Other IT Projects

Geographic dispersion of team members and the locus of project management are just two of the important risk factors affecting Internet-based projects. Project managers must also deal with the problem of being separated from their users, both geographically and temporally, which complicates the process of verifying work and seeking advice. Outsourced projects are likely to become Internet-based for convenience of the outsourcing company and the contractors. Even the software to support Internet-based project management might be outsourced from firms like iVenturi, a joint venture between Dow Chemical and Accenture (Gilbert, 2000). With such outsourced project management applications, the user interface of the support software becomes an important issue as well. Finally, security of project data is in jeopardy both during transmission and in storage at dispersed sites.

Locus of Project Management

An important decision that must be made at the outset of an Internet-based project is the physical location of the project manager. From the standpoint of the project manager, preference would be given to remaining in his or her current location. In the case of outsourced projects, however, the project manager might be a significant distance from the client, greatly complicating coordination during the project. The client might prefer that the project manager be located close to them, but then the manager would be separated from the bulk of the project team members, again causing coordination and supervision problems.

Even in the case of in-house projects, the project manager and team members may be working at different locations. The emphasis is on fast development, so geographic separation demands reliable, fast communications. Ken van Heel, a business manager in Dow Chemical's e-business group says, "Everyone's trying to get products to market faster. They need fast communication about hitting milestones and resolving problems" (Gilbert, 2000, 159).

Detached Users

When the project manager and team are geographically separated from the users, the gulf between them is even wider than that experienced in traditional software development projects. It becomes much more difficult to involve users in the regular decisions that must be made in the course of development, both in design and in coding. To use project support software or groupware for this purpose requires user training, adding upfront time and cost to the project. Users are also less likely to devote full attention to project team requests because of their lack of physical presence and the pressures of their day-to-day jobs.

Outsourced Development

Many of the problems that accompany outsourced development have already been addressed in traditional software project risk management (Schmidt et al., 2001). When the primary means of coordination and workflow management is through the Internet, however, a special burden is placed on the outsourcing firm to keep track of the contractor. Much as the users lose touch with the project, management is less likely to attend to project oversight—out of sight, out of mind. The first hint of problems with the project will inevitably have to come from the project manager, who might be reluctant to communicate these problems without the specific approval of contractor management.

The only alternative to this scenario is for client management to receive training in the use of the project support software and to use the software on a regular basis to track project activity. This should lead to a "shadow" project manager within the client organization. This alternative introduces more complications, such as the irritant of "being watched" and the possibility of conflict between the contract project manager and the overseer.

User Interface

The user interface of any software is one of the most important design considerations. Because of the special needs of Internet-based project management, practically all software supporting the project comes from vendors not associated with the users or the developers. Thus, the user interface of the project support software is not specifically designed for their purposes. Practically all firms offering such software, including iVenturi, Agile, Oracle, and SAP, are designing "one size fits all" products (Gilbert, 2000). With a widening user base, including the client users, client management, project manager, and project team members, the special needs of each type of user are less likely to be met in the user interface. This could lead to confusion, frustration, and less than perfect communication.

Security

Considering the vehicle for communication, geographic dispersion, and affiliation of stakeholders, Internet-based projects are especially vulnerable to security problems. Every use of coordination software exposes sensitive project data to compromise. Although high-level encryption is supported by modern Web browsers, not all project management software and groupware support such data protection. E-mail is also often sent unencrypted, so that routine communications among team members or between the team and users might not receive any protection at all.

With geographic dispersion of team members, the number of locations where sensitive data is stored is much greater than with traditional projects. Paper files, computer files, diagrams, audio and video recordings, software simulations and prototypes, screen images, and so forth are likely to be generated to facilitate communication among developers and between the developers and users. With the additional storage locations, the job of managing security of project data becomes complicated.

If the project is outsourced, the project participants belong to different organizations. The affiliation of those holding project data can get in the way of effective security management, from the standpoint of both physical surveillance and the enforcement of security policies. Attention to security issues can vary significantly from one organization to the next, so client management and the project manager might not agree on proper procedures and policies.

Sources of Risk

As a first step in formulating strategies to deal with risks, it is important to understand the sources of risk (Schmidt et al., 2001). There are two reasons for this. First, by recognizing the source of a risk, the project manager can take action directed at the source to improve the chance of a positive outcome, much as a doctor treats an infection rather than the fever that is a symptom of the infection. Second, some risk sources are perceived by managers to be beyond their control, which leads them to inactive management of the associated risks. By understanding the sources, project managers can take special care to attend to risks they might otherwise overlook.

Users

One factor that can plague Internet-based projects is lack of trust on the part of the users. Because the development team has only a virtual presence, users never develop the kind of relationship and the level of trust that is achievable when the team and the users are located together. The physical separation of users from the team also contributes to inadequate user involvement, historically one of the more serious risks to project success (Barki, Rivard, & Talbot, 1993; Schmidt et al., 2001). Because users do not have convenient access to the assembled team, roles and responsibilities can become muddled, leading to a further decline in user involvement. This is the phenomenon called "the detached user." There is also a strong possibility that users will develop expectations that are not in line with the expectations of the project team. Delivering software that does not match user expectations is another frequent cause of project failure (Schmidt et al., 2001).

General Management

The same lack of trust that can arise among users can also be found in general management. Poor professional relationships can lead to poor coordination of project activities or disinterest in the success of the project (Ash, 1998). The geographic and temporal separation of project team, users, and management also makes it difficult to maintain control over project dependencies with any external agents involved in the project.

As sponsor of the project, general management plays a special role in project success. The most important risk factor in software development projects is top management commitment to the project (Keil et al., 1998). If general management staff members lose focus of their stake in the project, their commitment becomes questionable.

Environment

Because project participants are not meeting face-to-face as in traditional projects, cues that might reveal to them

any changes in the business or political environment will be missed. For example, a change in upper-level leadership might result in changes in business strategy or organizational culture. The project as originally designed might not support the new strategic direction. The project also might continue along lines that are no longer politically acceptable. Even worse, the delivered system is at risk of poor alignment with the organizational culture, so when it is delivered it might not be used.

Technology

Technological risk is not new to software project management. Half of Boehm's (1989) top 10 risks were related to technology. But over the years, as reliability of computing technology and sophistication of the user interface have increased, the perception of risk associated with technology has decreased dramatically (Schmidt et al., 2001). In Internet-based projects, however, there is an increased risk associated with the use of technology to support the actual development process. Most of the technology designed to support Internet-based project management is less than three years old. Although there have been many studies of the use of groupware to support system development, using this technology across organizations in complex projects is not well understood. The majority of project managers have only limited experience and training with the new software and are only just beginning to understand the organizational dynamics of Internet-based projects.

Project Management

A different kind of project manager might be required for Internet-based projects. Internet-based projects tend to be more complex and under increased time pressure (Gilbert, 2000; Highsmith, 2000). Complexity arises not in the software development tasks, but in the coordination of activities with geographically and temporally dispersed team members, client management, and users. Time pressure arises through two phenomena. First, the users expect faster development because the use of Internet technology appears speedier than other means of communication. This is actually an unrealistic expectation (Ash, 1998). Rapidly communicating a requirement does not mean that the code satisfying the requirement will be written faster. Second, the startup time for an Internet-based project is longer because of the need for a period of training and orientation in use of the supporting software before actual work can proceed. So team members are under pressure to perform at a rate beyond their capability.

The Internet-based project is not just a traditional project on steroids. Because of the different ways of handling relationships with users and client management, the different supporting software, and the different way of managing the workflow, the skills demanded of the project manager are radically altered from those of traditional projects. Adequate techniques for managing Internet-based projects have not been developed or taught, and new methodologies especially designed for Internet-based projects are still being tested. Project management still has the potential to be a significant source of risks.

Project Team

If the project manager is not well prepared, how can the project team be prepared? Changes in personnel or staffing levels can have a wrenching effect on traditional projects. With Internet-based projects, such changes are more difficult to handle because the team members are often in different locales. Thus, bringing a new team member up to speed on the project is difficult. Also, because team members outside the organization are utilized, key personnel support resources may be difficult to access or nonexistent. Team members can become disheartened or disgruntled.

Because the project team is often pulled together using members at a distance from each other, their interpersonal relationships are tenuous at best. They lack the cohesiveness of a team that works together with daily "face time." Also, because they are remote from the project manager, there is an increased risk that other duties and activities may draw part of their time or attention, slowing project execution (Ash, 1998).

STRATEGIES FOR RISK MANAGEMENT

As mentioned in the introduction to this chapter, risk management is divided into two stages: assessing risk and controlling risk. The most difficult step in the first stage is identifying the risks that pertain to the project at hand. Historically, two methods have been favored for this task. The checklist method makes use of a list of potential risks. The brainstorming method relies on interactive input from experienced project managers and team members.

Once the appropriate risks have been identified, the project manager must take these risks into account when making project plans. Countermeasures should be devised in advance to minimize the damage of a negative outcome or to take advantage of a favorable outcome in the event one of the anticipated risks is triggered by events. The choice of countermeasures depends on an understanding of the risks in terms of a general classification. By recognizing the type of risk to be dealt with, specific strategies can be chosen that target the source of the risk. There are two principle means of countering risk. First, the manager may plan a project in such a way as to minimize the occurrence of certain risks. Second, when such planning is not possible, the manager might prepare specific actions to be taken in the event that a particular risk is triggered. Another approach might be to seek ways to hedge against particular risks that cannot be resolved by action (Kumar, 2002). For more complicated eventualities, the manager might develop scenarios that help the team understand what to do in the event certain circumstances occur (Ahn & Skudlark, 2002).

Recognizing Risks

To identify the risks associated with a particular project, the project manager must first be able to recognize risks. That is, the manager needs to know what the typical risks are. Each of the methods for risk identification has strengths and weaknesses in this regard. When choosing a method, the manager must take these strengths and

weaknesses into account. In this chapter, the two most popular methods are described. Charette (1989) describes other methods that might be used.

The Checklist

When managers lack a sufficiently broad experience with project management, they do not fully trust their own recognition of potential risks. In such case, a checklist of potential risks is essential. The checklist is useful for experienced managers as well, because it serves as a memory aid. Checklists are often compiled from "lessons learned" gleaned from analysis at project closeout. One problem with such checklists is cultural blindness to issues that may threaten a project. Schmidt et al. (2001) suggested that a more comprehensive checklist can help overcome cultural biases that may hinder the recognition of key risks.

There are two problems with the checklist approach. First, such a list may not be complete. Second, some risks on the list may have been overcome by events, making the list outdated. In the first case, the manager might overlook important risk factors, and in the second case the manager might spend valuable time preparing for a contingency that is not a risk at all.

For some years, Boehm's (1989) top 10 list was used as a starting point for risk analysis. By the mid-1990s, several independently compiled checklists were available. An examination of those checklists combined produces 33 distinct risk factors. A study that unified and expanded the existing checklists presented a list of 53 risk factors (Schmidt et al., 2001). Table 1 lists the 11 software project risks deemed most important by the international panel of experienced project managers polled in that study. Considering the currency of this unified list, it is not likely to suffer from the second weakness discussed here, but this list was compiled based on the experiences of project managers with traditional software projects. For that reason, it suffers from incompleteness when considering Internet-based project risk.

Table 1 Top Eleven Risks

RANK	RISKS
1	Lack of top management commitment to the project
2	Failure to gain user commitment
2	Misunderstanding the requirements
4	Lack of adequate user involvement
5	Lack of required knowledge or skills in project personnel
6	Lack of frozen requirements
7	Changing scope or objectives
8	Introduction of new technology
9	Failure to manage end user expectations
10	Insufficient or inappropriate staffing
11	Conflict between user departments

From "Identifying software project risks: An international Delphi study," by R. Schmidt, K. Lyytinen, M. Keil, and P. Cule, 2001, *Journal of Management Information Systems, 17*, 5–36. Copyright © M. E. Sharpe, Inc., 2001. Reprinted by permission.

The discussion of Internet-based project risks in the preceding section suggests several new risks that should be added to any checklist a project manager might use. It also suggests that the likelihood of some previously known risks' occurrence should be increased. There is no guarantee that all eventualities have been taken into account, however. The project manager needs more than just a checklist to prepare for risk management.

Brainstorming

The other major approach to risk identification is brainstorming. This method allows for the recognition of new and unusual risks that might be uniquely associated with a given project. Groupware provides an electronic forum in which brainstorming can be conducted across a wide group of participants regardless of their geographic or temporal separation. Special tools are available for capturing, classifying, and winnowing suggested risks to obtain as comprehensive a list as possible.

This method relies completely on the combined experience of the participants in the brainstorming session. If the participants lack sufficient experience, it is not likely that they will have encountered all the common risk factors, let alone less common risks that might apply to the project at hand. This is especially true for Internet-based projects, because most project managers have not had any experience with this form of project management, and even the seasoned Internet managers' expertise cannot compare with the long experience of traditional project managers.

General Principles

Thus, it is not likely that a single method of risk identification will suffice for any given project. The recommended approach is to supplement the checklist with a brainstorming session to identify potential risks not found on the checklist. The evaluation period immediately after brainstorming can help identify listed risk factors that are not pertinent as well.

After identifying the risks, the project manager must evaluate each risk in terms of its likely impact on the project, as well as the likelihood of its occurrence. This evaluation should be done in coordination with all the stakeholders in the project. These stakeholders will have a better appreciation of the expected outcomes for some risk factors. For others, the project manager should review the records of past projects of a similar nature in the organization. With this information in hand, the manager can assess the relative expected outcomes of the various risks. Once this is done, the more serious threats to the project will have been identified. Next, the manager will be ready to consider countermeasures for these risks and prioritization of action based on the relative expected outcomes.

Planning for Risk Countermeasures

Controlling risks is not a simple matter. Most managers would like to manage risks through elimination of the risk. This can be done by removing or countering any factors that might threaten the project. An examination of the lists of risk factors given by Schmidt et al. (2001) and found

in the previous section of this chapter reveals that these risks cannot be controlled in this manner. For example, the project manager cannot secure a 100% guarantee that all of the team members will stay with the project through completion. Even iron-clad contracts cannot counteract sickness, injury, or death.

Most managers take a rather simple approach to risk planning. They consider risks in three general categories: those that come from a source within their control, those that are beyond their control, and those that are a shared responsibility (Schmidt et al., 2001). Their behavior toward each category is based on their desire to manage the expected outcome of the risk down to zero. For factors beyond their control, managers are apt to take no action whatsoever (March & Shapira, 1987). For those within their control, managers have confidence in their ability to handle each risk effectively. Managers spend most of their time and effort on the risk factors that have a source that is a shared responsibility. This categorization of risks suffers from one serious flaw: It does not consider the probability of the occurrence of a given risk.

For example, the project manager would probably view the exit of a team member from the project as a factor beyond his or her control. Even though managers spend considerable effort on retention of team personnel, little or no effort would be expended to prepare for the eventuality that a team member would leave. Turnover in the IT field is high, however: In IT projects it runs to nearly 30% (Keil et al., 1998). Considering the budgetary and political impact of a serious delay in project completion, the expected outcome of a team member's exit is high. In their survey of American managers, Schmidt et al. (2001) found that this risk factor did not even make the list of risk factors to be ranked for potential action.

The manager must be aware of such biases and prepare plans accordingly. An important technique in this regard is to use a classification scheme that is free of bias. Once properly classified, the manager can choose the appropriate strategies for dealing with each class of risks.

Classifying Risks

There are numerous ways of classifying risks. One simple method is to look at the sources of the risks. Schmidt et al. (2001) identified 14 risk sources and sorted their 53 risk factors according to these sources. Dealing with risks in this manner immediately suggests some actions that can be taken to control the sources of the risks. It also provides the manager with an indication of where to direct action.

Keil et al. (1998) suggested a four-quadrant approach to classifying risks (Figure 1). They use two dimensions, the level of expected outcome and the level of control the manager has over the source of the risk, to map a 2×2 grid. Each quadrant of the grid then suggests some general approaches to controlling the risks associated with the quadrant.

Cule et al. (2000) took a similar approach but further suggested that appropriate coping behavior for the four quadrants should be effective for any risks in those quadrants. In such case, it would not be necessary to identify and assess every individual risk. By treating each class of risks rather than individual risks, the risk management process is greatly simplified.

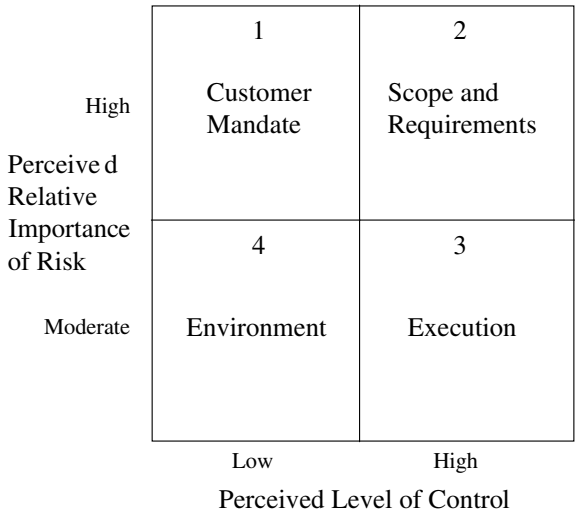


Figure 1: Risk Classification Framework. From Keil, et al. (1998). ©1998 ACM, Inc. Reprinted by permission.

Regardless of the approach taken to cope with risk, the classification of risk factors into four quadrants has considerable merit. It is obvious that no project manager could possibly attend to 53 or more risks, and it is doubtful that full attention could be paid to 14 sources of risks. Although a few large projects have used risk mitigation teams to cope with large numbers of risk factors, the vast majority of projects are too small to afford such an approach. By grouping the sources according to the scheme devised by Keil et al. (1998), the project manager can devise coherent strategies to counter the more serious risks facing the project.

It is important to understand that a specific risk factor might be assigned to a different quadrant from project to project, depending on its expected outcome and the assignment of authority. Consequently, the coping strategy chosen will shift from case to case for the same risk factor.

Choosing Appropriate Strategies

The highest priority group of risks to consider is risk factors that have a large negative expected outcome and are outside the direct control of the project manager. For example, if users and general managers develop a lack of trust in the project team, they may reject the team's work, leading to project cancellation or significant, costly delays. Coping strategies for such risks require relationship management, trust-building, and political skills (Keil et al., 1998). Because this is a high-threat category of risks, project managers must be prepared to take on these responsibilities.

But beyond the nature of these skills, the project manager must choose a specifically targeted approach to mitigate the anticipated risk. In the case of detached users and management, more rigorous scheduling and intensive performance tracking for reviews are necessary. A detailed communication plan should be in place to help mitigate this risk.

The remaining risk factors with a large negative expected outcome are directly under the control of the project manager. These risks largely can be controlled by

the project manager but do require skillful interfacing with the user or customer. Effective processes for managing change and ambiguity are needed. The real danger associated with risks in this category is that project managers may overestimate their own abilities or fail to realize their own limitations in dealing with the risks. With Internet-based projects, the lack of experience among project managers is a serious problem. If project managers assume the same practices they have used for traditional projects will suffice, it is likely that some of these risk factors will not be treated properly.

For risks that have a small negative expected outcome and fall under the direct control of the project manager, more routine project management techniques would be appropriate. The emphasis should be on internal and external reviews with the aim of keeping a project "on track." Considering the low threat these risks hold for the project, it would be more likely for the manager to try to treat these risks with a blanket approach as advocated by Cule et al. (2000). Direct attention to each of these risks would be disproportionate to their relative importance to the project.

The final class of risks includes those over which the project manager has little or no control and also have a small negative expected outcome. The primary problem with this class of risks is that because they are not viewed as being terribly important, the project manager is less likely to expend any political capital in getting anyone else to attend to the risks. Sometimes the small negative expected outcome is due to the low likelihood of occurrence of a risk. If such a risk has a large magnitude, then when it hits a project, its effects can be significant and dangerous. It is extremely difficult to plan for these risks. Because of their rare occurrence, there is a dearth of experience in dealing with them and it is impractical to engage in elaborate preventive measures to mitigate the risks. Contingency planning, including concepts and tactics associated with disaster planning, is the most sensible strategy for dealing with this class of risks. Scenario building is another approach for developing plans to deal with these risks should they occur (Phelps, Chan, & Kapsalis, 2001).

CONCLUSION

In this chapter, the possibility of new risks associated with Internet-based project management was explored. Existing knowledge about software project risk was summarized, and the concept was extended to cover specific situations typical of Internet-based projects. Four important concepts were emphasized. First, all risks can be traced to a source, and action taken to manage risks should be directed at the sources of risk. Second, identification of project risks is a key process in risk management. A hybrid approach using checklists of known risks combined with brainstorming to adjust the list to a specific project provides the best approach to identifying risks. Third, rather than classify risks according to source, the risks can be classified using a 2×2 grid that compares level of expected outcome with degree of control over the outcomes. Fourth, project managers can manage most risks with some simple coping behaviors. Some risks (those that pose the most serious threats to the project), however, require specific, targeted treatment.

GLOSSARY

Expected outcome The product of the magnitude of an outcome and the probability of its occurrence.

Groupware Software specifically designed to support group interaction, including brainstorming, collaborative work, and consensus building.

Internet-based project Any project that is principally or wholly managed through the use of Internet tools, rather than through face-to-face interaction.

Risk An event with an uncertain outcome.

Risk factor Any factor that could affect the outcome of a risk.

Workflow management The active oversight and direction of a series of tasks by any group of actors working in concert toward a common goal.

CROSS REFERENCES

See *E-business ROI Simulations; Electronic Commerce and Electronic Business; Prototyping; Return on Investment Analysis for E-business Projects*.

REFERENCES

Ahn, J., & Skudlark, A. (2002). Managing risks in a new telecommunications service development process through a scenario planning approach. *Journal of Information Technology 17,* 103–118.

Arrow, K. (1965). *Aspects of the theory of risk-bearing*. Helsinki: Yrjö Jahnssonin säätio.

Ash, T. (1998). Seven reasons why Internet projects fail. *UNIX Review's Performance Computing, 16,* 15–16.

Barki, H., Rivard, S., & Talbot, J. (1993). Toward an assessment of software development risk. *Journal of Management Information Systems, 10,* 203–225.

Benett, G. (1998). Working together, apart: The Web as project infrastructure. Retrieved October 12, 2002, from http://www.intranetjournal.com/features/idm0398-pm1.shtml

Boehm, B. (1989). *Software risk management tutorial.* Washington, DC: IEEE Computer Society Press.

Bumbo, N., & Coleman, D. (2000). Collaboration for distributed project management. Retrieved October 17, 2002, from http://www.collaborate.com/mem/case_studies/casestudy3.php3

Charette, R. (1989). *Software engineering risk analysis and management*. New York: McGraw-Hill.

Collaborative Strategies. (2000). Electronic collaboration on the Internet and intranets. Retrieved October 17, 2002, from http://www.collaborate.com/mem/white_papers/intranet.php3

Cule, P., Schmidt, R., Lyytinen, K., & Keil, M. (2000). Strategies for heading off IS project failures. *Information Systems Management, 17–2,* 65–73.

Gibbs, W. (1994). Software's chronic crisis. *Scientific American, 271,* 86–95.

Gilbert, A. (2000). Online project management planned. *Informationweek, 808,* 159.

Highsmith, J. (2000). *Adaptive software development: a collaborative approach to managing complex systems*. New York: Dorset House.

Keil, M., Cule, P., Lyytinen, K., Schmidt, R. (1998). A framework for identifying software project risks. *Communications of the ACM, 41,* 76–83.

Kumar, R. (2002). Managing risks in IT projects: An options perspective. *Information & Management 40,* 63–74.

Lyytinen, K., & Hirschheim, R. (1987). Information systems failures—A survey and classification of the empirical literature. In Zorkoczy, P. I. (Ed.), *Oxford surveys in information technology* (Vol. 4, pp. 257–309). Oxford, UK: Oxford University Press.

Lyytinen, K., Mathiassen, L., & Ropponen, J. (1998). Attention shaping and software risk: A categorical analysis of four classical risk management approaches. *Information Systems Research, 9,* 233–255.

March, J., & Shapira, Z. (1987). Managerial perspectives on risk and risk taking. *Management Science, 33,* 1404–1418.

Phelps, J., Chan, C., & Kapsalis, S. (2001). Does scenario planning affect performance? Two explanatory studies. *Journal of Business Research, 51,* 223–232.

Schmidt, R., Lyytinen, K., Keil, M., & Cule, P. (2001). Identifying software project risks: An international Delphi study. *Journal of Management Information Systems, 17,* 5–36.

van Genuchten, M. (1991). Why is software late? An empirical study of the reason for delay in software development. *IEEE Transactions on Software Engineering, 17,* 582–590.

Walkerden, F., & Jeffery, R. (1997). Software cost estimation: A review of models, processes, and practice. In *Advances in Computers* (Vol. 44, pp. 62–94). San Diego, CA: Academic Press.

Rule-Based and Expert Systems

Robert J. Schalkoff, *Clemson University*

INTRODUCTION

One of the most widely used models of knowledge representation and manipulation is the production system. Rule-based systems may be thought of as a subclass of production systems. Production systems are conceptually simple and, when implemented as shells in "canned" form, may be developed with a minimum of specialized programming. Examples of such system shells are OPS5, CLIPS, and Corvid (see "OSHA Compliance Monitoring and Advising" later in the chapter).

The term "expert system," although seemingly a catchall in current jargon, is used to indicate a subset of production systems that are restricted to specific task domains. Many expert systems (ESs) are implemented using the rule-based paradigm, in which knowledge is encoded in "if–then" form.

The explosive growth of the Internet has created a large application environment for rule-based expert systems. Many of these systems are packaged inside so-called intelligent agents. Applications include recommender systems for e-commerce, online portfolio selection, network monitoring, and expert advisors for OSHA (U.S. Occupational Safety and Health Administration) compliance monitoring. In addition, there are efforts to standardize the communication of business-related rules across the Web. These are described more fully in Selected Internet Applications at the end of the chapter.

THE PRODUCTION SYSTEM PARADIGM AND RULE-BASED SYSTEMS (RBS)

Overview

Figure 1 shows a somewhat simplistic and generic rule-based production system (Schalkoff, 1990) consisting of the following:

1. A database of information or knowledge (e.g., facts).
2. A set of productions (e.g., rules) that modify the existing database and whose applicability is conditioned on the current database.
3. A control mechanism or rule interpreter that determines the applicability of the rules in the context of the current database and selects the most appropriate rule(s) through a process known as conflict resolution.

Productions in production systems are specified by a set of condition–action pairs. Specification of conditions in the form of "if" statements and actions via "then" yields the familiar rule-based system representation. Production systems may also be thought of as a subset of pattern directed systems, systems in which production applications are driven by input (or initial) data patterns. Production system operation is based on a "production cycle" of searching for applicable rules, selecting a particular applicable rule, using (firing) the rule, and repeating the cycle. Firing of a rule usually results in some modification (e.g., addition or deletion of facts) of the current database.

Rule-based systems provide a natural means to express situation–action heuristics in problem solving. They are also a natural means to express observed human behavior. Thus, a rule-based paradigm is a natural choice for expert system implementation. Rule-based expert systems are typically based on if–then (implication)–based representations of knowledge obtained from expert query and applied in narrowly defined problem domains.

Following is an example of a rule from the DENDRAL (Lindsay, Buchanan, Feigenbaum, & Lederberg, 1980) expert system. DENDRAL is one of the oldest ES, having been developed in 1964 (in the programming language LISP). A typical production in DENDRAL (Harmon/King, 1985) is the following:

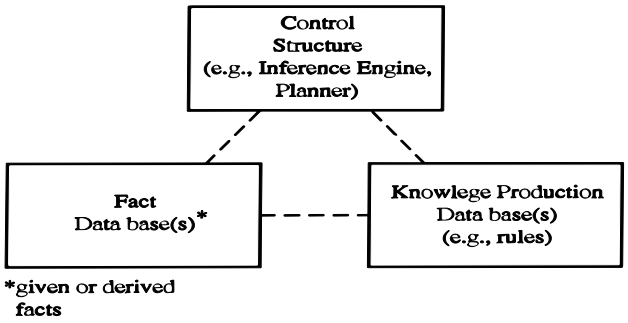


Figure 1: Basic production system architecture.

If
the molecule spectrum has two peaks at masses
$x1$ and $x2$ such that
$x1 + x2 = M + 28$ AND
$x1 - 28$ is a "high" peak AND
$x2 - 28$ is a "high" peak AND
at least one of the $x1$ or $x2$ is high,
Then
The molecule contains a ketone group.

Another sample rule, taken from the online technical support example in the "Web-Based Technical Support" section is this:

if "symptom" = "Label Designer hangs on converting
databases dialog"
and ("BackTrack Version" <= "4.10")
and "second edition" = "yes"
and (operating_system = "Windows 98 Second Edition"
or operating_system = "Windows 98")
then problem = "Win 98se"

Figure 2 shows a more realistic rule-based system architecture, which takes into account practical concerns such as user interfacing and structuring and partitioning of the fact and rule databases.

Features of Rule-Based Production Systems

Advantages

Characteristic of systems employing productions in rule form are the following generally positive features:

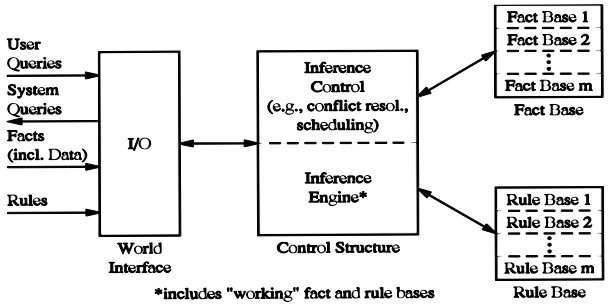


Figure 2: More complete rule-based production system structure. Includes structured rule and facts databases and user interface.

1. Expressibility (rules allow the expression of knowledge in a form understood by human experts and at the same time sufficiently quantitative for symbolic manipulation by the machine).
2. Ease of modification of the database (facts and or rules may be added or removed); in addition, the development and implementation of systems that are highly modular (e.g., see Figure 2) is straightforward.
3. Ease of exploring the current knowledge base contained in the system (i.e., the encoding of information is in a readable form).
4. Flexibility of processing (the inference mechanism(s) may be chosen to suit the problem).
5. Ease in following the inference mechanism (the order in which rules were employed may be recorded and traced for an "explanation" of the system's conclusions).
6. Standardization in terms of a knowledge representation and inference approach.
7. Availability of off-the-shelf software for implementation.

Disadvantages

Disadvantages of productions in rule form include the following:

1. Inability to predict system behavior for a given problem. (This characterizes many AI computations.) In other words, there may be many solutions, no solution or a unique solution, and the only way to establish this is via IE search.
2. Inability to force a specific production sequence compared with imperative programming. (Note some expert system shells allow manipulation of inference engine parameters to accomplish this.)
3. Lack of suitability for all applications.
4. Lack of ability to implement directly "deep reasoning" and "common sense." Note that this is a general representational issue, not an inherent shortcoming of the rule-based production system paradigm.

THEORETICAL AND COMPUTATIONAL ASPECTS OF RULE-BASED SYSTEMS

The Logical Basis of Rule-Based Inference

Generation of new facts or verication of a goal fact set in a rule-based production system proceeds by linking the "then" parts of rules to the "if" parts of other rules and proceeding until the goal statement is proven to be true or no new facts may be produced. This process is called chaining and is accomplished under the control of the inference engine (IE).

Consider a simplistic rule of the (abstract) form:

$$\text{If } p \text{ then } q,$$

where p is denoted the rule antecedent and q is denoted the consequent. More formally, p and q are statements in logic, connected by implication (the symbol $\rightarrow$ denotes the implication connective), and this is normally written

as

$$p \rightarrow q. \tag{1}$$

Modus ponens (MP) is one logical basis for chaining and is based on the axiom

$$\{(p \cap (p \rightarrow q)) \rightarrow q\} = T. \tag{2}$$

Thus, given a rule (which itself is assumed to be TRUE) of the form $p \rightarrow q$, to prove $q(= T)$, we must verify $p(= T)$ in the database. Typically, Equation 2 is shown with the following notation:

$$\begin{array}{r} p \\ \underline{p \rightarrow q} \\ q \end{array} \tag{3}$$

Rules typically involve variables to enable a more general representation. For example, a rule indicating "X is the grandparent of C" in the form of Equation 1 might look like this:

$$[(\text{parent } XY) \cap (\text{parent } YZ)] \rightarrow (\text{grandparent } XZ). \tag{4}$$

In this example, notice the use of variables that appear multiple times (i.e., are shared) in the antecedent (i.e., variable Y) and those that are shared across the antecedent and consequent (i.e., variables X and Z). Assignment of a value to a variable is called binding. Given a database of clauses of the form (parent tom sally) and so on, it is necessary to find a consistent set of bindings for variables X, Y, and Z to use the rule in Equation 4. Such a set of bindings is called a unifying substitution. When rules involve variables, as in the preceding example, MP is revised as follows:

$$\begin{array}{rl} p^1 & \text{(statement(s) without variables)} \\ p \rightarrow q & \text{(rule with variables)} \\ \dfrac{(\exists\theta)(p^1 = p\theta)}{q^1 = q\theta} & \text{(unifying substitution).} \end{array} \tag{5}$$

THE CONCEPT OF CHAINING AND INFERENCE DIRECTIONS

Chaining is an important concept in the implementation of a rule-based system. It is typical that the consequents of some rules are the antecedents to others (using unifying substitutions, where necessary). These links form potential "chains," which are logical links or paths through the system rulebase. A forward chaining (or antecedent-driven) production system attempts to form chains from the initial fact base to a database containing the goal. A backward or consequent-driven paradigm attempts to form (conditionally) chains backward from a goal database to the initial facts database. Hybrid strategies involve both forward and backward chaining. Whether a production system is implemented through a forward, backward, or hybrid chaining paradigm, the IE searches for one or more paths through the problem state space and may therefore explore a large number of redundant or unsuccessful paths in the process. A good IE design uses all available a priori information (such as properties of commutativity and decomposability, if applicable), to avoid needless or unproductive searching.

Potential Complexities in Chaining

Referring to Equation 2, rule-based inference becomes complicated in a number of realistic situations, including the following:

1. There exist many rules to choose from, that is, the database contains multiple rules of the form

$$\begin{array}{lll} p_1 & \rightarrow q & (= T) \\ p_2 & \rightarrow q & (= T) \\ & & \vdots \\ p_n & \rightarrow q & (= T). \end{array} \tag{6}$$

   Thus, we have the initial problem of choosing a rule.
2. The antecedent p in a rule is not a simple statement, but a compound expression in logic, for example,

$$p = p_1 \cap p_2 \cap p_3 \ldots \cap p_n. \tag{7}$$

   Thus, it is necessary for the IE to verify $p_1 \cap p_2 \cap p_3 \ldots \cap p_n$.
3. Assuming (shared) variables are involved, there are multiple bindings that would satisfy all antecedents in the structure of the antecedent in Equation 7.

All of these circumstances may occur in realistic problems.

THE INFERENCE ENGINE (IE)

The heart of a rule-based production system is a database (facts and rules) and the corresponding inference mechanism or inference "engine" that manipulates this database. As noted, the inference engine (IE) is responsible for

1. Selection of relevant rules (or sets of rules), which may pertain to a specific reasoning scenario that the control mechanism determines should be focused on.
2. Determining the suitability of a given rule, using the criteria listed here, given the current database.
3. Execution or firing of the rule(s) and subsequent modification of the database.
4. Determination if the overall goal has been satisfied.

In practice, the control strategies used to implement conflict resolution (rule selection) are diverse. They may range from tests as simple as "fire the first rule found to be applicable" to "fire the rule that (according to some heuristic) gets the system closest to the desired goal state."

It is entirely possible for the IE to be formed from a production system in which meta-rules (rules about rules) are used to guide the selection of production system rules.

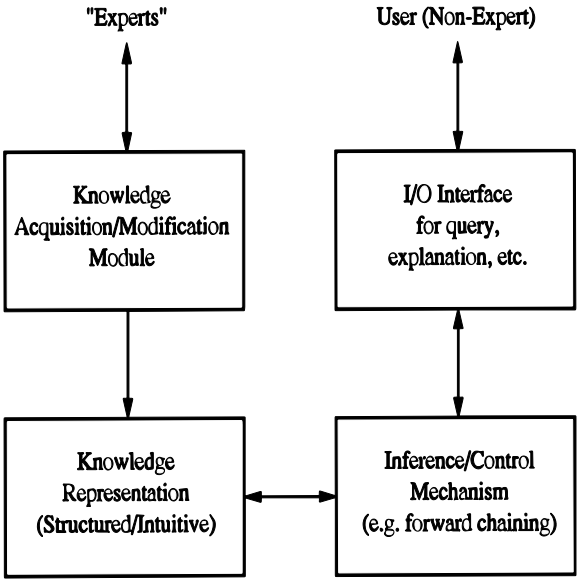


Figure 3: Overall expert system structure indicating both development phase (knowledge engineering) as well as system interface to nonexpert.

Although perhaps enabling a more sophisticated rule selection mechanism, this approach necessitates the design of another meta-rule-based IE.

RULE-BASED EXPERT SYSTEMS DEVELOPMENT

The term "expert system," although seemingly a catchall, is used to indicate subset of production systems that are restricted to specific task domains: Expert systems are programs that attempt to emulate the behavior of human experts, usually confined to a specific field. Expert system shells and knowledge acquisition systems have been developed using disparate approaches to knowledge representation and manipulation as well as user interfacing. A sample ES structure is shown in Figure 3.

The following are typical attributes of an ES:

1. Knowledge is usually represented in declarative form to enable easy reading and modification. Most ESs use if–then structures for representation; thus, rule-based ESs predominate.
2. There is usually a clear structure to the knowledge representation (this excludes neural expert systems).
3. There is a clear distinction between the knowledge representation and the control or manipulation mechanism.
4. A significant user I/O (input/output) interface is provided to allow query, advice, explanation, and interaction with the ES.
5. A user knowledge acquisition or knowledge modification module is often provided for extension of the ES.

The Appeal of Expert Systems

The development of ES is motivated by a number of factors, including the following:

1. Expert-level knowledge is a scarce and expensive resource.
2. ESs make expert behavior available to a large audience. This is especially true of those implemented using the Internet.
3. The integration of the expertise of several experts may lead to ESs that outperform any single expert.
4. ESs are not motivated to call in sick, leave a company for better working conditions, or demand huge salaries (although their development and maintenance costs are often substantial).

The potential for expert systems is enormous. Declining development costs have led to numerous efforts in developing both small, easily modified ES, as well as large systems. The quantification coding of human insight, compassion, motivation, guessing ability, and learning capabilities is still an elusive goal, however. Often the ES design process requires a minimum of new technology and a large amount of engineering judgment.

Expert System Examples

Many ESs have been developed and are currently in operation. An early example of a commercially successful systems is XCON from Digital Equipment, which configured computer systems. XCON was written in OPS5, a rule-based programming language. Other historically significant examples include MYCIN (Buchanan & Shortliffe, 1984) and CADUCEUS, used for medical diagnosis, and PROSPECTOR, which guided geological prospecting. When PROSPECTOR found a molybdenum deposit worth $100 million (U.S.), this application gained respect.

Expert System Challenges and Limitations

One might expect the performance of expert systems, which could tirelessly and exhaustively consider every possibility associated with a problem, to outperform humans in a spectrum of applications. This is currently not the case. ES developers have discovered that knowledge acquisition can be slow, expensive, and iterative. Furthermore, systems tend to be "brittle" in the sense that slight modifications in the application lead to unacceptable deviations in ES performance. It is not incidental that a human spends approximately 12 years past the age of 5 (or so) in formal schooling. Notwithstanding the possible lack of efficiency in this process, a significant amount of both information and experience (which is perhaps not as easily quantifiable) is gained over this time interval. In addition, most perceived experts have a considerable amount of additional informal and formal education past this point. Thus, we should not be surprised at even the practical difficulty of representing expert behavior.

Expert System Development

The first questions an expert system developer must ask are the following: Are bona fide experts available whose performance is significantly better than that of amateurs? Can their expertise be automated? and Does it make practical and economic sense to develop an ES?

The development of expert systems proceeds with the consulting of an expert (or group of experts) with the aim of developing a manipulable knowledge base. This is often referred to as knowledge engineering (KE). The first phase of the process consists of the formation of a database of domain specific knowledge. In the expert interrogation process, the formulation of "good" questions is paramount. Fortunately, experts often articulate problem-solving methodologies in terms of if–then structures. Moreover, an expert may volunteer rationalizations of the resulting rules (i.e., "I conclude this because . . . "). This type of explanatory production is also desirable.

The development of an expert system is almost always an iterative task, involving the cycle of expert query, database formation development of the inference strategy, verification of system performance, and so on. The gap between the concept of an ES and a finished, delivered product may be enormous. The necessary application-specific selection of a reasoning structure, interviewing of experts, development of a prototype, refinement, user training, and documentation may take several years.

One of the most important aspects of ES development is verification of system operation. A set of test cases is developed and used by both the ES and the human experts. When responses differ, modifications to the system and perhaps additional expert consultation are required.

The ability of an ES to provide the user with an explanation is also important. An expert system response such as "patient has disease x" is probably insufficient, even if correct, because no explanation of the inference process is provided. An explanation may be as simple as indicating the sequence of rules used or as complicated as indicating all possible inference paths considered and the logic that leads to the most appropriate response or conclusion.

Reasoning With Uncertainty

Unfortunately, knowledge may seldom be put into the rule-based if–then form without some concern for concepts such as impreciseness, ambiguity, and uncertainty. Although several techniques (e.g., fuzzy systems) are treated more fully elsewhere, the incorporation of measures of uncertainty in the representation as well as the inference (manipulation) strategy leads to more realistic ES implementations. A number of significant research efforts have attempted to incorporate uncertainty in inference techniques. These are the following:

1. Numerical approaches, which attempt to associate a quantitative measure of confidence (or certainty) with the truth value of facts. This includes fuzzy set and probabilistic approaches. For example, "certainty factors" have been used to quantify uncertainty in expert reasoning. In the expert system MYCIN, a confidence scale of [−1,1] was used to represent the range of confidence associated with a particular fact or assertion (conclusion). A value of −1 indicates total lack of confidence (i.e., complete confidence the assertion is FALSE), whereas a measure of 1 represents complete certainty the assertion is TRUE. If, after exhausting all search possibilities, the cumulative confidence measure associated with a hypothesis is in the interval [−0.2, 0.2], the hypothesis is regarded as unconfirmed. Of course, this is an empirically determined range that is subject to modification or alternate interpretation in a particular application.
2. Symbolic approaches incorporate uncertainty but in a less numerically quantitative manner. Examples are linguistic extensions to the connectives of classical logic that allow statements using "perhaps," "may be," and so on. An example rule is the following:

```
If  symptom is believed to be spots,
then  diagnosis may be measles.
```

3. Another approach is the use of multivalued logic (Bolc & Borowik, 1992).

Rule-Based Expert Systems and Intelligent Agents

The notion of agents brings together a number of technologies and research areas, including artificial intelligence, software engineering, robotics, and distributed computing. Agents are a powerful, natural metaphor for conceptualizing, designing, and implementing many complex, distributed applications. A quantitative definition is "an agent is an encapsulated computer system that is situated in some environment, and that is capable of flexible, autonomous action on behalf of its user (or owner) in that environment in order to meet prespecified design objectives."

The Internet is arguably one of the most complex, changing, and unpredictable environments with which software designers must deal work. At the same time, Internet applications are arguably one of the most important areas from both technical and economic perspectives. The Internet may offer one of the best opportunities for the agent paradigm.

The current trend is that of multiple agents; a multiagent system is one that consists of a number of agents that interact with each another as well as the environment. Characteristics of an agent include the following:

1. An agent is a problem-solving entity with well-defined boundaries and interfaces.
2. An agent is embedded in a particular environment.
3. An agent is designed to achieve specific objectives.
4. An agent is autonomous.
5. An agent is flexible and displays (context-dependent) problem-solving behavior. In other words, the agent is reactive.

One connection between agents and expert and rule-based systems is straightforward: Expert or intelligent agents may be implemented using a rule-based paradigm. Sample applications are shown in Katz (2002) and Grosof (1995).

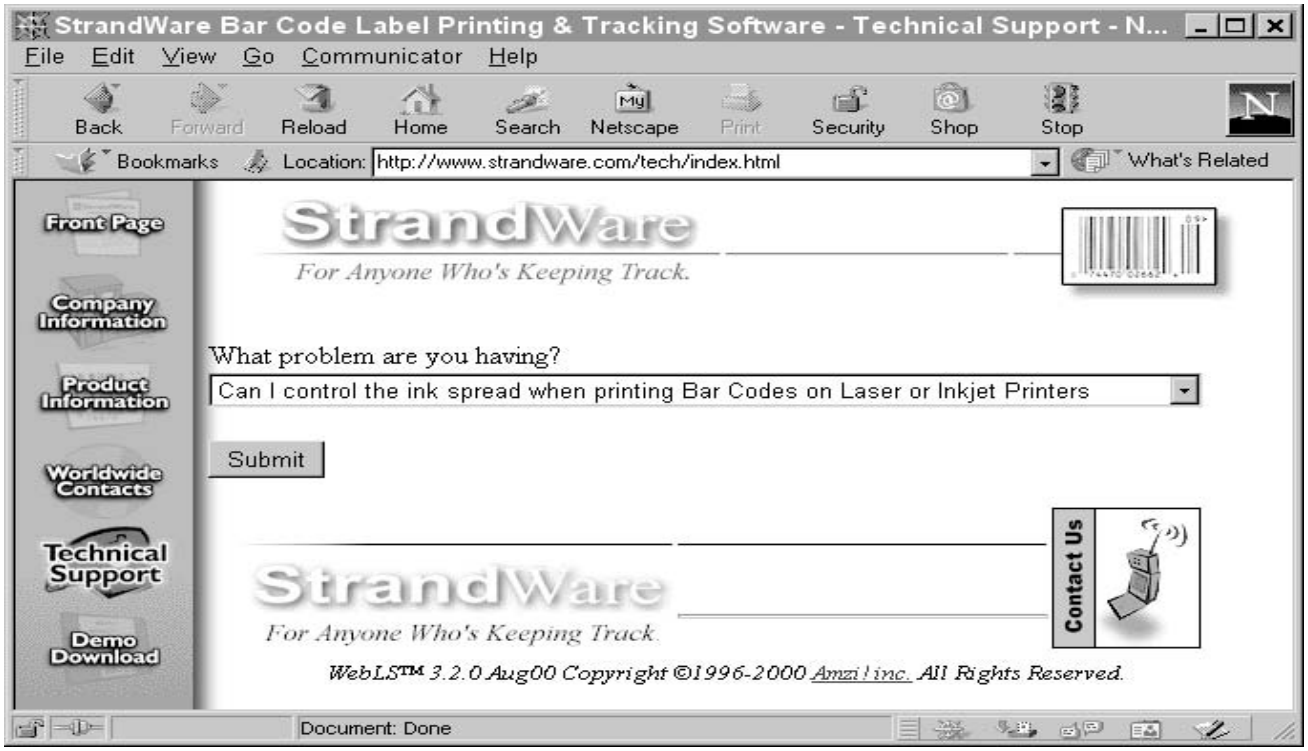


Figure 4: Strandware help session.

SELECTED INTERNET APPLICATIONS

Web-Based Technical Support

Numerous examples of Internet-based technical support exist. Software companies, in particular, observed early on that many product technical support issues revolve around problems that recur. Web-based technical support for these problems provides a number of benefits for both the company and its customers. For example, a certain percentage of customer support issues may be solved autonomously. The customer gets a quick resolution of the problem, and technical support resources are conserved for the more challenging problems. In addition, autonomous Web-based technical support serves as a filter for human technical support. When a customer does not get a resolution to the problem using the online expert system, human technical support may be invoked. An additional benefit is that preliminary problem information has already been acquired by the autonomous system and thus the duration required for the session is shortened. Finally, the autonomous system may gather all the data from many Web-based support sessions and thus provide a comprehensive database of previously encountered problems and corresponding solutions.

An example of this approach is Strandware, which develops bar-code label design and data collection software for industrial and business tracking applications (http://www.strandware.com/). (A sample rule was shown earlier.) The rule-based automated technical support system is implemented using Amzi's WebLS product. WebLS is a simple ES tool, designed specifically for deploying expert systems on the Web.

When the customer initiates a support session, he or she is asked for information regarding the product and problem. The system uses this initial information to filter the rules that might apply and gathers additional information as needed in its search for a problem resolution. A sample screen is shown in Figure 4.

The Strandware ES is composed of questions, rules, and answers. An initial goal starts the system searching through the rules. The antecedents of some rules in this system are facts with values that can be obtained by querying the user. The following are two such facts related to diagnosing a 16-bit ODBC (open database connectivity) problem:

```
question("second edition", [
     prompt = $Is it Second
        Edition or have you
               installed the Y2K
                 update?<BR>$,
     ask =yes_no,
     askAfter = ["symptom"]
]).
question("BackTrack Version", [
     prompt = $What version of
       BackTrack are you
               running?<BR>$,
     ask = field,
     askAfter = ['symptom']
]).
```

Details of the Strandware system are provided at http://www.amzi.com/customers/strandware.htm and in Hicks (2000).

Electronic Commerce: Recommender Systems

E-commerce provides many opportunities for advanced automation, including gathering and using information on potential customers, buying trends, and product purchases. This information may be used to suggest future purchases to previous customers. A knowledge-based recommender system (KBRS) generates purchase recommendations by consulting a product information knowledge base and then reasoning what products will best satisfy perceived user requirements (Burke, 1999). One technique for building a recommender system is a knowledge-based approach called the PersonalLogic system. This system helps users make decisions on wide range of product choices. The system first acquires user requirements in a particular product domain (e.g., automobiles) and then consults its knowledge base to find suitable items that satisfy the users' requirements. An automobile shopper could provide requirements such as automobile type, size, features, and price range, and the system would then search its knowledge base for automobiles that best satisfy these requirements.

Online Portfolio Selection

PORSEL (PORtfolio SELector) (Zargham & Mohammad, 1999) is a system for analysis and selection of stocks. PORSEL allows for fundamental analysis and uses an expert system. PORSEL is a Web-based client–server system, with the stock analysis programs and an associated information database residing on the server. The Web allows remote access to the system.

PORSEL uses a fuzzy rule-based system to perform fundamental analysis. Fundamental analysis primarily uses information about a company to predict future movements of the company's stock. This is the approach articulated by a number of famous investors, thus a large body of knowledge is available for use in creating rules. Most rules are derived from well-known rules, such as Graham's rule:

> If a stock has a price-to-earnings ratio of less than 40% of the stock's highest price-to-earnings ratio over the past 5 years, then the stock's rating is good.

PORSEL has shown excellent performance when compared with the Standard and Poor's 500. Sample results are shown in Table 1. (Note that these results represent retrospective, not prospective, performance.) Here, all shares were purchased at the beginning of the year, held for the entire year, and sold at the year's end. PORSEL then selected 20 new stocks for the the next year. "Equal proportion" means that the same amount was invested in each of the selected stocks, whereas "Variable proportion" indicates that PORSEL also optimized the relative amount invested in each selected stock.

Table 1 Sample PORSEL Performance

YEAR	EQUAL PROPORTION	VARIABLE PROPORTION	S&P 500 INDEX
1989	55.5	58.8	31.49
1990	73.06	89.5	−3.17
1991	109.3	226.5	30.55
1992	29.1	−4.0	7.67
1993	567.4	573.5	9.99
1994	38.1	44.4	−1.50
Average	145.41	164.78	12.5

S&P = Standard and Poor.

Network Monitoring

ExperNet (Vlahavas, 2002) is a multiagent system for monitoring computer networks, detecting problems, and diagnosing the source of problems. Each agent in the system is responsible for managing a portion of the network (e.g., a single agent manages each subnet). Using multiple agents has many advantages, including fault tolerance and a reduction in the amount of monitoring information transmitted over the network.

Each agent has a modular structure. The "Device" knowledge base system comprises the expert system shell in which rules are implemented. Device is implemented on top of CS-Prolog II. CS-Prolog II uses an extension of HNMS network management software, called HNMS+, to acquire information about the network. The system also uses a computer monitoring program called BigBrother to gather information about the computer to which the agent is attached. This information is used to infer information about the network's performance. Device provides many interesting features, such as support for multiple rule types (deductive, production, and event-driven rules) and object orientation.

OSHA Compliance Monitoring and Advising

OSHA compliance monitoring, for many companies, is a complex, time-consuming, and important issue. Rather than providing a static database, OSHA, using the Corvid ES development tools from EXSYS (http://www.exsys.com/exsys.html), developed eTools for this purpose. These tools are available at http://www.osha.gov/dts/osta/oshasoft/. According to OSHA,

> eTools are "stand-alone," interactive, Web-based training tools on occupational safety and health topics. They are highly illustrated and utilize graphical menus as well as expert system modules. These modules enable the user to answer questions, and receive reliable advice on how OSHA regulations apply to their work site.

Corvid allows the development of interactive experts that deliver individualized decision-making knowledge from a Web page or as stand-alone systems. The Exsys CORVID developer interface provides two ways to view the underlying system logic through Logic Blocks and individual rules. The Rule View window shows the full text of each rule in English, making the rule simple to read and interpret. A wide range of interfaces and controls are available for how questions are asked of the system user and for integration with other software.

CORVID Logic Blocks may be integrated with forward or backward chaining. Furthermore, Logic Blocks are built and maintained in a visual development environment. The underlying knowledge representation in the block is an if–then rule.

CommonRules

Although not strictly a simple ES, the IBM CommonRules framework is highly related to both e-commerce and rule-based systems. In 1999, IBM developed an application called CommonRules 1.0: Business Rules for E-Commerce. Its goal was to make a program available that would allow businesses to communicate their "business policy rules about pricing, promotions, customer service provisions for refunds and cancellation, ordering lead time, and other contractual terms & conditions, to a customer application/agent." An overview may be found at http://www.research.ibm.com/rules/commonrules-overview.html. According to IBM,

> CommonRules 3.1 is a rule-based framework for developing rule-based applications with emphasis on maximum separation of business logic and data, conflict handling, and interoperability of rules.
>
> . . . it provides a platform that enables the rapid development of rule-based applications through its situated rule engine via dynamic and real-time connection with business objects. CommonRules can be integrated with existing applications at a specific point of interest, or it can be used to create applications composed only of rules. (IBM Common Rules, n.d.)

Each business involved in e-commerce probably implements its own set of rules, implemented using many different languages. Because these representations are different, communicating business policy to customers or other businesses becomes challenging. The purpose of CommonRules is to eliminate this difficulty by using a semantically rich rule language called CLP (Courteous Logic Program). CLP incorporates many rule sets and a rule-based computational model. XML (extensible markup language) is the string format used for CLP, which makes it a natural format for the Web.

CommonRules 1.0 was developed in Java and runs on the user's computer. It accesses the Internet when invoked. When the user wishes to look at a seller's available business rules, the user downloads these rules and runs them through CommonRules to output a file with a set of user-readable rules. This eliminates the customer's needing to have the same rule representation of the seller to view its business policies. The current version of CommonRules is 3.0 and can be found at http://alphaworks.ibm.com/aw.nsf/download/commonrules.

CONCLUSION

Many internet applications may be enhanced with the perception of intelligent behavior. This "intelligent" enhancement may be achieved using a rule-based or agent-based computational paradigm. Examples include e-commerce, diagnosis, and compliance monitoring.

GLOSSARY

Agent Encapsulated and task-focused software structure.
Antecedent The "if" part of a rule.
Certainty factor One method to incorporate uncertainty in the inference process.
Conflict resolution Process of choosing which rule or rules to use in the inference process.
Consequent The "then" part of a rule.
Expert system Software intended to emulate human expertise.
Heuristics "Rules of thumb" intended, but not guaranteed, to aid in the solution of a problem.
Inference engine Production system controlling software.
Knowledge representation The paradigm chosen to encode knowledge (e.g., rules and facts).
Modus ponens A logical basis for rule-based inference.
Multivalued logic A logical system with more than two truth values.
Production (rule-based) system A conceptual and computational paradigm useful for building intelligent and expert systems.
Recommender system Software that tracks previous purchases and extrapolates to future recommendations.
Rule If–Then structure used to represent expertise or knowledge.

CROSS REFERENCES

See *Intelligent Agents; Machine Learning and Data Mining on the Web.*

REFERENCES

Bolc, L., & Borowik, P. (1992). *Many valued logics—1. Theoretical foundations.* Berlin: Springer.

Buchanan, B. G., & Shortliffe, E. H. (1984). *Rule-based expert systems: The MYCIN experiments of the Stanford Heuristic Programming Project.* Reading, MA: Addison-Wesley.

Burke, R. (2000). Knowledge-based recommender systems. In A. Kent (Ed.), *Encyclopedia of Library and Information Systems* (Vol. 69, Suppl. 32). New York: Marcel Dekker.

Grosof, B., Levine, D., Chan, H., Parris C., & Auerbach J. (1995). Reusable architecture for embedding rule-based intelligence in information agents. *Proceedings of the Workshop on Intelligent Information Agents, ACM Conference on Information and Knowledge Management (CIKM-95).* New York: ACM Press.

Harmon, P., & King, D. (1985). *Expert systems—artificial intelligence in business.* New York: Wiley.

Hicks, R. (2000, January/February). New Trends in ES development and implementation. *PC AI Magazine, 14,* 37.

IBM CommonRules (n.d.). Retrieved April 14, 2003, from http://www.research.ibm.com/rules/commonrules-overview.html

Jackson, P. (1999). *Introduction to expert systems*. (3rd ed.). Reading, MA: Addison-Wesley.

Katz, E. (2002). A multiple rule engine-based agent control architecture. In *Proceedings of the 6th IEEE International Conference on Intelligent Engineering Systems*. Piscataway, NJ: IEEE Press.

Lindsay, R. K., Buchanan, B. G., Feigenbaum, E. A., & Lederberg, J. (1980). *Application of artificial intelligence for chemistry: The DENDRAL project*. New York: McGraw-Hill.

Schalkoff, R. J. (1990). *Artificial intelligence, an engineering approach*. New York: McGraw-Hill.

Vlahavas, I. (2002). ExperNet: An intelligent multiagent system for WAN management. *IEEE Intelligent Systems, 17,* 62–72.

Wooldridge, M., & Jennings, N. R. (1995). Intelligent agents: theory and practice. *The Knowledge Engineering Review, 10,* 115–152.

Zargham, M., & Mohammad, S. (1999). A web-based information system for stock selection and evaluation. In *International Conference on Advanced Issues of E-Commerce and Web-Based Information Systems, WECWIS* (pp. 81–83). Piscataway, NJ: IEEE Press.

FURTHER READING

Retrieved April 14, 2003, from http://www.exsys.com/demomain.html

Retrieved April 14, 2003, from http://www.exsys.com/whitepaper.html (CORVID Technical Specifications)

Retrieved April 14, 2003, from http://www.osha.gov/dts/osta/oshasoft/ (OSHA eTools page)

Retrieved April 14, 2003, from http://www.amzi.com (Amzi Software Tools)

S

Secure Electronic Transactions (SET)

Mark S. Merkow, *E-commerce Guide*

INTRODUCTION TO SECURE ELECTRONIC TRANSACTIONS (SET)

Credit card theft on the Internet has reached epidemic proportions, and everyone who handles credit card numbers and expiration dates clearly needs to understand that the handling is akin to *toxic chemical handling* and mandates the utmost of care and diligence. The risks of theft and misuse of credit card data by thieves and nefarious users who target the databases and systems that store and maintain the data are too great to ignore or treat casually.

Daily reports of security breaches, extortion, identity theft, and general havoc continue to dog e-commerce and drive away large proportions of the buying public. To partially answer these concerns, the banking associations—Visa and Mastercard—jointly issued Secure Electronic Transactions (SET) as a specification to implement the business services needed for worldwide processing of credit, debit, and charge card transactions over open channels like the Internet.

SET opens the doors to e-commerce but comes with a steep price both in time and in dollars to implement. SET is complex—so complex that possible future use of the banking standard remains an open question. Unlike most other efforts aimed at secure e-commerce, SET mandates the involvement of all its participants—buyers, suppliers, card processors, and back-end bank system operators. SET compliance requires onerous efforts on everyone's part. In 2002, many industry observers and experts would say that SET is dead with the selling public still satisfied to use secure sockets layer (SSL), but perhaps SET is only hibernating, awaiting an elusive market catalyst.

Background

Early in the 1990s, banks were refusing to accept or process charges originating on the Internet and required merchants who wanted to sell their merchandise online to use existing infrastructures (dial-up, etc.) for charge authorizations; point-of-sale transactions, phoned-in requests for charge authorizations, and follow-on batch processing activities. These banks, led by pressures on two sides—merchants and consumers—began pressuring the Visa and MasterCard Associations to develop secure standards for using credit cards over any insecure channel, such as the Internet.

Visa and Microsoft responded with one standard they released in September 1995. The *Secure Transaction Technology* (STT) specification was posted to the Visa Web site for download by interested parties. At the same time, Microsoft announced that it would develop STT implementation tools for Windows 95 and Windows NT that could be licensed by developers. Tools for other desktop platforms would be developed by Spyglass Technology, which was behind what is known in 2002 as Microsoft's Internet Explorer software.

Meanwhile, MasterCard and its allies, Netscape, IBM, Cybercash, and GTE (now Baltimore Technologies), had developed the *Secure Electronic Payment Protocol* (SEPP) as a proposed specification and posted it to the

MasterCard Web site for a public comment period. MasterCard had hoped that SEPP could be in use for Internet transactions as early as April 1996.

Incompatible Payment Card Standards

STT and SEPP generated such heated debate and finger pointing between the two opposing factions that the entire industry was at odds. Both sides claimed their standards were defined with "openness in mind" and was designed in cooperation with the Internet standards-setting bodies, the W3 Consortium, the Internet Engineering Task Force (IETF), CommerceNet, and the Financial Services Technology Consortium.

Industry and financial services observers at the time believed that STT and SEPP were similar, yet different enough to render them incompatible. This meant that separate implementation efforts were required, and all parties desiring to support Visa and MasterCard products needed separate processing systems to meet the unique requirements of each protocol.

Not So Different After All

In fact, STT and SEPP both attempted to achieve the same objectives, but did so from different directions. These objectives included

Changing Internet-based credit card transactions from the risky *card-not-present* scenario (such as mail order/telephone order transactions) to the less risky *card-present* situation (such as retail shopping and eateries) to reduce the chances of fraud and increase the potential to lower the fees that merchants must pay for maintaining merchant accounts.

Requiring all parties in a transaction (customer, merchant, credit card processor, or bank) to possess digital certificates that establish their identities and their authority to conduct transactions.

Requiring that public-key certification agencies (Certificate Authorities) manage the certificate distribution processes on behalf of the card associations or member banks.

Using industry-standard public-key cryptography techniques, as developed by Ron Rivest, Adi Shamir, and Leonard Adelman (RSA Data Security).

Encrypting only credit card numbers and transactional data rather than encrypting the entire browser and shopping sessions.

Concealing credit card data from all merchants to prevent merchant-initiated fraud and reduce the risk of operating a merchant commerce server.

Enabling use of *any* type of credit card product, regardless of issuer. The card associations however, reserved the right to specify that only *their* protocols be permitted for transactions with their cards.

By the end of 1995, banks that issued both Visa and MasterCard were up in arms against the attempt to force two separate standards that accomplished the same task. The banks persisted and finally forced Visa and MasterCard to work together on a *single* standard, because supporting two separate ones appeared unachievable.

In February 1996, an announcement rocked the Internet community:

> Visa & MasterCard Combine Secure Specifications for Card Transactions on the Internet Into One Standard.

SET Consortium Established

According to the agreement, Visa and MasterCard, along with GTE, IBM, Microsoft, Netscape Communications Corp., SAIC, Terisa Systems, Verisign, and RSA Data Security formed the SET Consortium. Its goal was to resolve the differences and conflicts between STT and SEPP and develop a new unified standard.

The development of SET arose not so much from a spirit of mutual cooperation as from the intervention of major banks that saw the industry giants, Visa and MasterCard, heading in separate, nonintersecting directions. Obviously, for any pragmatic solution to the problems of electronic commerce, a single, standard approach, both flexible and platform-independent, was needed.

SET is designed to eliminate most problems of credit card usage and data management on the Internet, adding the following elements:

Message authentication to assure entities involved in a credit card transaction that they are dealing with whom they think they are dealing

Data integrity to prevent spoofed messages

Confidentiality to prevent eavesdropping

The first version of SET, based on the work of the SET Consortium, was released to software developers in draft form on June 24, 1996. The draft release, embodied in three separate electronic books, was intended to be used for testing and to elicit comments from outside experts. These books contained preliminary specifications sufficient for developers to build components that would "bolt on" to existing cardholder browsers, merchant commerce servers, and financial institution credit authorization systems. SET Version 0.0 appeared as follows:

Book 1—The business description containing background information and processing flows. It was intended as a primer on software that interfaces with payment systems and employs public-key cryptography.

Book 2—The programmer's guide containing the technical specifications for the protocol, intended for use by software developers who wished to build cardholder and merchant software components.

Book 3—The formal protocol definition, intended for use by cryptographers analyzing SET's security aspects, writers producing programming guides for toolkits or components, and system programmers developing cryptographic and messaging primitives. The formal protocol is cast in Abstract Syntax Notation (ASN.1).

With the initial release, SET Version 0.0 was placed under change control with a January 1997 deadline for enhancement requests to Version 1.0, and March 1997 for proposed corrections to the testing version. On April 21,

1997, SET Version 0.2 was released to the public, containing requests for enhancements that satisfied the additional needs of non-Visa/MasterCard issuers, such as American Express, Japan Credit Bank (JCB), and Novus/Discover.

On May 31, 1997, SET Version 1.0 was released to the public.

Complying with the SET Standard

Because SET is an open and neutral protocol, in theory it is possible to purchase any implementation from anyone who offers it without concern for proprietary ingredients. To turn this theory into a reality, independent testing is required to ensure compliance as defined by the specification.

SET Version 1.0 was the baseline standard that developers used to build their systems, knowing that later versions will most certainly supersede it. As SET versions evolved, software needed to be tested and retested for compliance with the new standards.

Developer interpretation of the specification was at the root of the problem. For SET to ever succeed, it needed a single, unambiguous understanding among developers that eliminated the possibility of proprietary implementations of SET. That is one of the major tasks of Secure Electronic Transaction, LLP, or SETCo.

Compliance Testing and Certification

SETCo operates under the sponsorship of the card associations, but is independent of them. They assume the responsibility for SET's development, maintenance, evolution, and market acceptance, and they regulate the use of the SETmark for products that successfully pass a rigorous compliance testing program. SETCo also maintains a dispute resolution board that decides how to best handle disputes or questions regarding an implementation. The SET Compliance Administrator (SCA) serves the administrative functions for SETCo, evaluating test results submitted by software developers and maintaining SET testing tool suites.

Upon signing SETCo's Master License Agreement and paying the fees, developers are given testing tools and scripts. These testing tools run through various permutations of SET messages, monitor, and log the results, and assist in identifying noncompliance (if it occurs). Once a developer is satisfied that its product is compliant, it sends the test results back to SETCo for verification. After SETCo is satisfied, it permits the developer to use the SETmark on its software. Each compliant software component of SET will bear its own SETmark, attesting that the component itself passed the battery of tests. The SETmark, however, does not ensure that the component will work properly with other counterpart SET components build by other developers. SETCo does not offer end-to-end testing services, nor does SETCo offer interoperability testing services.

Baseline vs. Derivative Systems

SETCo established rules separating the products that must be tested and certified from the ones that need not be. Each unrelated operating system implementation is considered a baseline product and must undergo full testing with submission of test results to SETCo. Derivative products are adaptations of baseline products intended for use on similar operating systems. An example of a derivative is a port from Solaris to AIX. The derivative must be documented for SETCo's purposes, and testing is strongly encouraged, but test results need not be submitted and no additional fees are charged. Developers decide which products they consider the baselines and which ones the derivatives.

In addition to the initial testing of the baseline version, SETCo requires developers to retest their products every 6 months. In the re-evaluation process, retesting is performed using the latest testing scripts, with the results resubmitted to SETCo for evaluation. If recertification is not performed, the developer risks losing the use of the SETmark. Periodically, SETCo may decide to audit completed software compliance tests. Reasons for these audits include random selection to assure conformance and reports from the field indicating suspected failures or compromises of security. To remain in the compliance-testing program, developers must pay an annual fee to continue licensing SETCo's services.

Whereas SETCo oversees the testing that ensures direct compliance of distinct components, the interoperability testing between components and between products from different companies remains the domain of SET developers, who take on the work themselves.

Vendor Interoperability Testing

Recognizing that interoperability is a critical hurdle, most SET developers agreed that their products must be tested for interoperability before they were introduced into the marketplace. Therefore, IBM and VeriFone produced the "Interoperability Reference Guide for SET Version 1.0," based on IBM and VeriFone's experiences culled from the testing they performed. The document describes test scenarios, assumptions in use, configurations for environments, and other related information. RSA Data Security Inc. also developed interoperability documentation to assist developers. The SET 1.0 Interoperability Test Plan defines a certificate infrastructure, data, and business scenarios that vendors may use to test their applications among themselves.

CREDIT CARD PROCESSING AND CORRESPONDING SET PHASES

To the uninitiated, credit card processing may appear straightforward and simple, but the complexity involved is hidden under the covers. To help understand SET within the context of its intended use, the following is a quick explanation of who is involved and what happens when a credit card charge is made and the goods are delivered.

Roles in Card Processing

A cardholder is the user of a credit card that was applied for and received from an Issuer Bank.

A merchant is an accepting destination for a credit card charge

An acquiring bank manages merchant relationships and accepts charge receipts from merchants as deposits into Merchant Accounts.

A card association (e.g., Visa and Mastercard) dictates the conditions under which its branded cards may be used and provides the network and processing to permit the three primary constituents (cardholder, merchant, and acquiring bank) to transact business.

Basic Credit Card Schemes

There are two major approaches to credit card schemes—closed loops and open loops. In a closed loop system, the issuer and the acquirer are the same organization—they manage both the cardholder and merchant relationships. Examples of closed loop systems include Discover (Novus), American Express, Japan Credit Bank (JCB), and Diner's Club (operated by Citibank). In an open loop system, the issuer of a credit card may or may not be the same as the acquiring bank. Because the Visa and Mastercard networks consist of well over 20,000 banks worldwide, there are a tremendous number of possible combinations (4×10^8) of issued cards and acquirer processors for any given charge transaction. For example, a cardholder holding a Visa card issued by Bank A may shop at a merchant who has a merchant account at Bank B. As the charge card is swiped on the merchant's point of sale (POS) terminal, a charge request is initiated and sent to Bank B (the acquiring bank), which places a charge authorization request on VisaNet. The Visa network then routes the request to Bank A to determine account status and sufficiency of credit for approving a new charge to the account. The response to this authorization request is an authorization response, containing an approve or decline status, along with a code for the merchant to use when the sale is completed (goods are shipped) and the merchant is ready to settle the charges.

SET provides the specifications for *request–response message pairs* that permits the parties involved to use open networks like the Internet to perform the same work that previously was performed using the private networks that the banks mandated for moving credit card information around. These message pairs offer the same business services that the private-network POS system offers, without the cost of dedicated network links and maintenance.

SET message pairs correspond to the following business services used for credit card processing:

Inquiry into charge transaction status

Payment processing

Authorization reversal when customers change their minds

Capture reversal when goods are returned for credit

Credit issuance when goods are returned for credit

Batch administration to settle charges and clear transactions

Certificate issuance for all entities

Certificate inquiries on pending certificate requests or status information

Error handling

SET Digital Certificate Management

SET certificate management and processing are in addition to any other transaction-based processing that takes occurs. The purpose of such processing is to ensure that certificates are current, are accurate, and are always ready for use when needed. SET cardholder certificates are constructed to mimic both the physical piece of plastic and the signature on the back of it. Merchant certificates ensure the transaction acquirer and the cardholders that they are dealing with a legitimate operator who is contractually obligated to the brand to remain honest. Charge processors and merchants are ensured that they are dealing with cardholders who have legitimate rights to use a brand product. Both merchants and cardholders are ensured that their transactions are seen and processed only by those charge processors who have legitimate rights to see and process them.

In some cases, a SET payment gateway is needed to validate SET digital certificates and preprocess authorization, capture, and settlement work. Payment gateways are operated by companies that perform the charge processing duties for merchants and banks. The terms *acquirer payment gateway* and *payment gateway* are synonyms. One example of a card processor in the U.S. is First Data Corporation in Omaha.

Because an SET merchant server takes the place of POS terminals, it needs to perform all the work POS terminals do, and then some. One significant benefit of using the Internet, rather than private networks and dial-up lines, is its flexible nature, which makes it possible to communicate freely. With Internet connections, it is possible to avoid some third-party work (with a resultant saving of their fees) by connecting directly to acquiring banks or card company payment gateways.

SET in Action During Charge Processing

SET is implemented as pairs of request and response messages that are enciphered using strong cryptography before being placed onto the public Internet to hide their contents to all but those intended to receive and process them.

In person, it is easy to check for a matching signature on a card or to ask a person for an ID. On the Internet, it is virtually impossible. Authentication thus can only occur through cryptography. SET uses a robust set of digital certificates to accomplish the identification and authentication activity. Each participant in a SET transaction requires a specific certificate or set of certificates that not only uniquely identify them, but also attest to their privileges as holders of payment cards or merchant accounts.

Before any transaction can take place, everyone involved needs one or more SET digital certificates. Without now looking specifically at how they are obtained, assume that the digital certificate issuance process has already occurred and everyone is prepared. Call this Phase 0.

Phase 0: All SET Software and Requisite Digital Certificates in Place

A cardholder of a Bank A credit card possesses a corresponding cardholder digital certificate and has installed the SET E-wallet software to operate with their Web browser. A merchant with an account at Bank B has installed the SET Merchant POS System and installed the requisite merchant digital certificates to enable it to operate. Bank B's payment gateway is up and running with

a payment gateway digital certificate for each credit card brand that it services.

Phase 1: The Shopping Experience

A cardholder visits a merchant's site, browses through the online catalog, and makes decisions about which goods to purchase for delivery.

Phase 2: Item Selections

As the cardholder selects goods to purchase, he or she adds the goods to the shopping cart that the merchant server uses, and the system dynamically tallies up the sale.

Phase 3: Check Out

Just as a shopper pushes his shopping cart to the cash register, the merchant server responds in kind when the cardholder proceeds to "Check Out." The shopping cart program adds up the costs of the items in it, adds sales tax, computes delivery and handling fees, and presents a list of selected items and the total back to the customer. When the cardholder is satisfied with the order, he or she proceeds to the payment selection phase.

Phase 4: Form of Payment Selection

With order totals still displayed on the screen, the consumer is given a choice of payment options. Assume the cardholder has selected a SET-enrolled Visa Card as the form of payment. *SET is now initiated*.

Phase 5: Payment Initiation Processing

By virtue of selecting a SET-enabled Visa payment, the merchant server sends a special message to the cardholder's browser, telling the e-wallet to *wake up*. This wakeup message tells the e-wallet to prompt the consumer to enter a secret password to unlock the wallet. The wakeup message also initiates the first SET payment processing message pair, called the Payment Initialization Request, which is generated and sent back to the merchant SET POS software (a component of the merchant server cash register functions). With a successful payment initialization response, the e-wallet then creates a SET purchase request message. This message has two components—a purchase order piece and a payment instructions piece. The merchant POS software can only read the purchase order. The payment instructions, containing information about the cardholder's Visa account, can only be deciphered and processed by the acquiring bank systems.

Phase 6: Payment Authorization Request

Along with the payment instructions, the merchant's SET POS software prepares a SET authorization request message intended for Bank B's payment gateway. The message contains the details about the amount of sale, the merchant account requesting it, and the previously created payment instructions component that the cardholder e-wallet software generated in Phase 5. When the authorization request is deciphered, the payment gateway creates a standard authorization request and places it on the bank's back-end private interchange network, which locates the cardholder's account at Bank A. With an approval code from Bank A to proceed with the sale, the payment gateway responds with a SET authorization response that tells the merchant's POS software to complete the sale. The

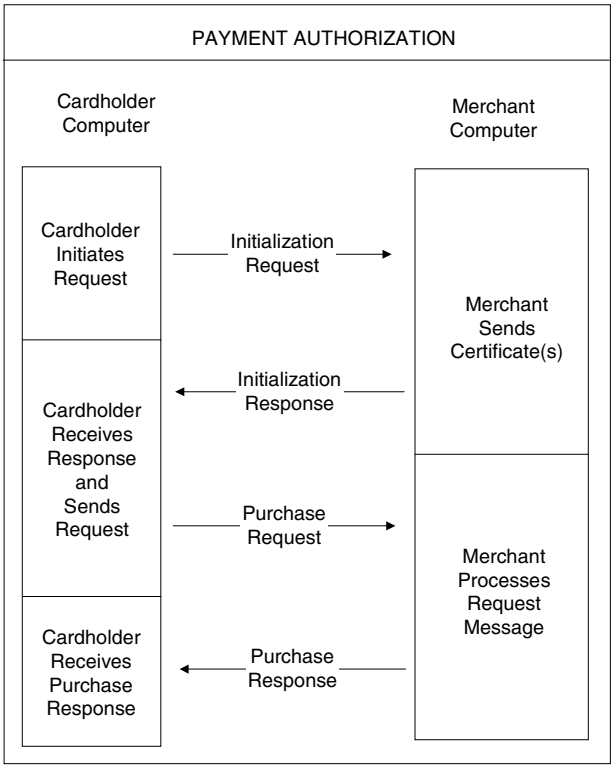


Figure 1: SET payment authorization flow between cardholder and merchant.

POS system then creates a purchase response message intended for the cardholder's e-wallet to confirm the sale and produce an electronic version of a receipt or record of charge.

Phase 7: Delivery of Goods

When the merchant POS system notifies the merchant's back office that a new order is ready for shipment, a shipping record is prepared and the merchant's POS system is informed of the business event via a data entry form.

Phase 8: Capture and Settlement

With the successful authorization code from Phase 6, the merchant's SET POS software received a capture record (SET calls these tokens). With the sale completed and the goods delivered, the POS software can initiate a capture request to finalize the sale with Bank B's payment gateway system. With each capture response, the settlement file grows, awaiting the merchant's decision to deposit these receipts into his or her merchant account at Bank B in exchange for funds transfer. Settlement file or batch processing is also carried-out via the Internet using SET's batch administration message pairs, designed specifically for those purposes.

See Figures 1 and 2 for illustrations of the flow for a SET payment authorization request/response message. Figure 1 illustrates the flow between cardholder and the merchant. Figure 2 illustrates the flow between the merchant and the payment gateway.

Although SET's actual processing work is identical to the work initiated via a POS terminal operating on a private network, SET makes it possible to use the Internet through its cryptography and message-passing

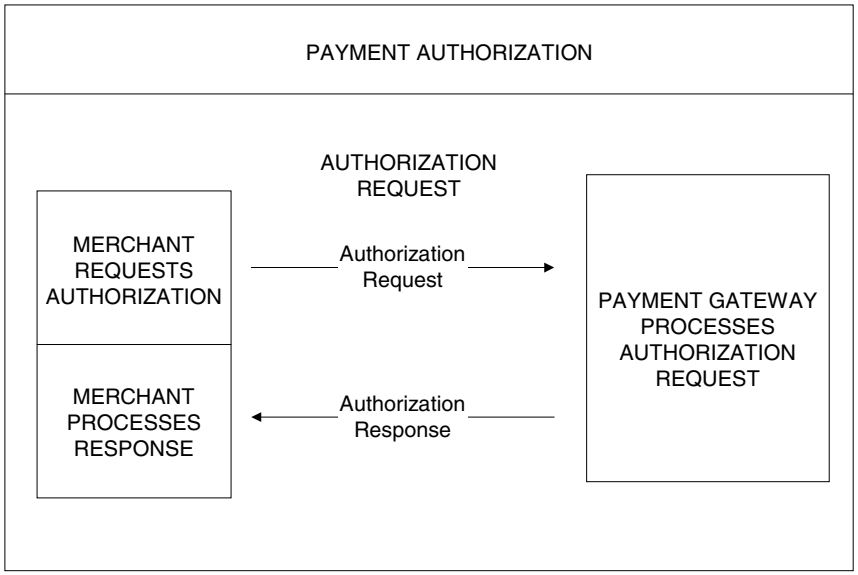


Figure 2: SET payment authorization flow between merchant and payment gateway.

mechanisms. It turns the *public* Internet into a *private* network that protects every SET message pair.

SET uses two forms of messages that relate to requests for and responses to processing between the cardholder and the merchant, and between the merchant and the acquirer payment cateway. There is never a direct link between the cardholder and the payment gateway—the merchant always serves as the message broker between the two parties.

Digital Certificates for SET

Digital certificates represent identity for all SET participants by binding a person's identity to a pair of electronic encryption keys that are later used to encrypt or sign digital information. A digital certificate helps to verify someone's electronically transmitted claim that he or she is who he or she claims to be and has the right to use the encryption keys. SET digital certificates prevent people from using stolen or fraudulent keys to impersonate other people. Used in conjunction with encryption, digital certificates provide a more complete security mechanism than simple ID and password mechanisms and SSL protections. The contents of a generic digital certificate may include the following:

Owner's public key.
Owner's name.
Expiration date of the public key.
Name of the certificate issuer.
Serial number of the certificate.
Digital signature over the entire certificate created by the certificate issuer (CA).

The most widely accepted format for digital certificates is defined by the CCITT X.509 international standard; thus, such certificates can be read or written by any application complying with X.509. SET's version of digital certificates is a special "flavor" designed exclusively for credit cards. SET *extends* the X.509 standard for e-commerce to permit its international presence without concern for export controls on encryption products or services.

See Figure 3 for a representation of a basic X.509 digital certificate.

Certifying SET Participants

SET mandates that all users obtain salient key-pairs in a secure manner that is impervious to attacks. Because the cracking of keys requires inordinate time and effort, would-be thieves typically will strike at the management and maintenance systems that store keys, rather than through cryptanalysis of the keys themselves.

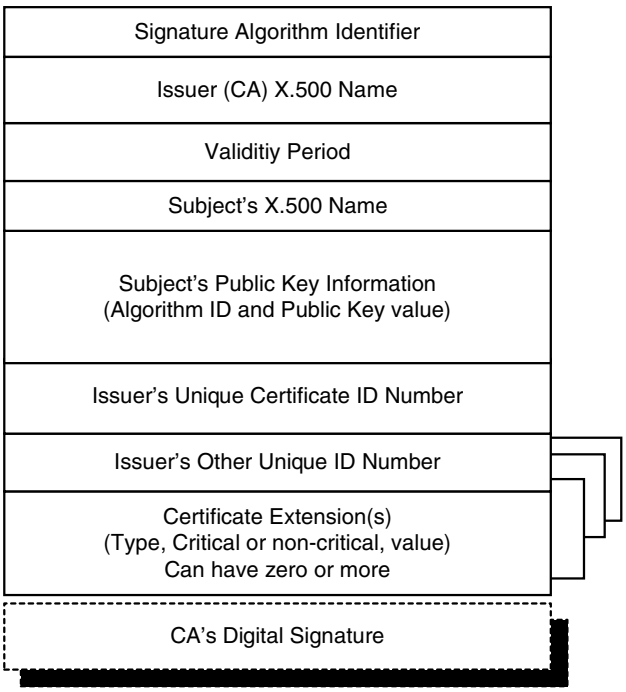


Figure 3: A basic X.509 digital certificate.

Table 1 Certificate Types Summary

Certificate Types	Digital Signature	Key Encryption	Certificate and CRL Signing
Cardholder	X		
Merchant	X	X	
Payment Gateway	X	X	
Cardholder Certificate Authority	X	X	X
Merchant Certificate Authority	X	X	X
Payment Gateway Certificate Authority	X	X	X
Brand Geo-Political Certificate Authority	X		X
Brand Certificate Authority			X
Root Certificate Authority			X

Digital certificates can aid in the effort to keep private keys secure while aiding in the dissemination of the related public keys (which together form the key pair). If a private key is disclosed (either by cryptanalysis or by theft), that fact needs to be shared so that the recipients of messages signed with the stolen key know to reject or disregard them. The mechanism that X.509 dictates and SET uses to determine if a private key has been compromised and reported is called the Certificate Revocation List, or CRL.

SET digital certificates attest to the binding of an end entity's public key to the end entity itself. Suppose Charles presents his payment instructions using his Visa Platinum Card from Bank A. When a merchant receives his message and subsequently forwards it for processing, both the merchant and the payment gateway can verify Charles' claim that the message and the certificate are his and no one else's. Because the private key tied to the Visa brand certificate was used to sign Charles' credit-card-based certificate, and Charles' message can only be decrypted using Charles' public key from the certificate, two things must be true; the message must have come from Charles, and upon successful certificate validation within the Visa brand tree of trust, the certificate must have been signed by the Visa brand certificate and no other.

In their basic forms, digital certificates contain the private-key holder's public key (half of the keypair), his or her name, the certificate's expiration date, a serial number, the name of the authority that issued the certificate, the policies under which certificate use is permissible, and any other information the issuer deems vital or useful. Most important, it contains the digital signature of the issuer. SET Certificates follow the ITU Recommendation X.509 for Version 3 certificates.

Summary of Certificate Types

SET's public key certificate hierarchy, or tree of trust, is an arrangement of certificate authorities (CAs) that implements the needs of each SET participant brand (Visa, Mastercard, etc.). These CAs hang off a root CA, operated by SETCo, which operates as SET's managing authority. The possible certificate types that may be present in the tree are summarized in Table 1.

SET APPEARS ON THE MARKET

With the release of Version 0.2 in 1997, developers began exploring how SET might be developed as a suite of products that covered the entire span of relationships mandated by SET. By 1998, new SET-compliant products appeared, such as the following:

Cardholder E-wallets

Integrated SET merchant POS systems and e-commerce systems

Payment gateway software for acquiring banks and card processors

Cardholder E-wallets

The first e-wallets appearing on the scene as SET-compliant immediately created an untenable situation for banks and cardholders. Banks do not tend to be very good at distributing software and software updates to millions of users. This fact, coupled with consumer skepticism of e-commerce as it existed in 1997 and the size of e-wallet download files, which ran up to 10MB downloaded over dial-up connections, led to downright uninterest and anger from cardholders.

Merchant POS Servers

The first suites of products intended to satisfy merchant requirements for SET could only be touched by those with the greatest patience and deepest pockets. Implementation would mean tens or hundreds of thousands of dollars in new hardware and software purchases in order to meet the stringent requirements for security. Banks did little to offer incentives to merchants to build SET compatibility into e-commerce, and at the same time, SSL-based security of credit card data over the Internet was thought *good enough*. Merchant uninterest in SET soon followed.

Payment Gateway Systems

A few pioneering banks began to develop their own payment gateway interfaces for SET, and SET pilot testing with willing merchants or simulated e-commerce systems using employees of the banks began to reveal problems in implementation and compatibility of implementations across software developers.

In the United States, NationsBank of Charlotte, North Carolina conducted the bank's first SET transaction in early 1998, demonstrating *some* interoperability between SET software providers. NationsBank worked closely with IBM, MasterCard, GlobeSet, and GTE to create a system for purchase of items from the MasterCard Emporium, an initial Web site built by the association to help consumers make small initial purchases and overcome their fears of shopping electronically.

Interoperability actually turned out to be more elusive than anyone had originally thought. For example, if an IBM CommercePOINT e-wallet failed to send properly formatted and decipherable messages to a GlobeSet POS system or vice versa, the IBM product would lock out customers or payment processors except for those who used the same software—creating an intolerable situation. Although any given SET suite may have worked flawlessly across all its own components, it was of no use if did not also work with components from other SET software developers. The situation was similar to the problem of a Web site owner insisting that visitors use one brand of browser to the exclusion of all other browsers. Merchants will never be certain which e-wallets shoppers use, nor can they know which payment gateway software their Acquirer bank uses.

By the middle of 1999, it looked like implementing SET—without adequate incentives from the issuers—was leading to a train wreck, and by 2000, several of the software providers for SET systems had sold off their assets and closed their doors forever.

With the ongoing pilot testing of SET Version 1.0, banks were finding that their consumers were experiencing agony and resentment when asked to deal with bank-branded electronic wallets. Banks were erroneously hoping their customers could perform these tasks:

Select SET-compliant e-wallets

Download them

Install them correctly on their PCs

Acquire the digital certificates for each credit card they wanted to use on-line

Always use SET for shopping and refuse to shop with merchants who were not SET-compliant.

Between the excessive downloading times for the huge e-wallet files (6MB+), encrypted transmissions that caused bandwidth-sapping operations, and unacceptable waiting times while transactions completed their processing, the testing banks and SET developers were forced to consider alternative software distribution approaches. Complicating the problems further, testing banks and SET developers assumed that merchants (and their webmasters) would become experts in banking systems for payment card processing. Initial implementations of POS components required near-ideal secure hosting environments, robust cryptographic processing facilities, and full understanding of banking and bank internetworks. As it turned out, only merchants with colossal patience could begin to meet these requirements.

Industry Attempts to Assuage SET User Concerns

GlobeSet, out of Austin, Texas, was the first to introduce the GlobeSet ServerPOS as an alternative to the GlobeSet POS System for merchant commerce servers. With ServerPOS, the merchant's acquiring bank or card service provider operates the POS component on its premises with a merchant storefront adapter that resides on the merchant's commerce server.

ServerPOS used a multiprotocol approach, supporting both SET and SSL, to encompass all online merchants with whom the acquiring bank had forged relationships. On the consumer front, GlobeSet also offered ServerWallet, using a similar scheme. With the GlobeSet ServerWallet, card issuer banks bear the responsibility for managing the SET consumer digital certificates (and public keys), transaction data related to purchases made with their cards, and the other SET-related security data. Consumers needed only to download a thin-client component (45 Kbytes in size) to start the communications with ServerWallet at the bank when a SET transaction was initiated.

Trials of ServerWallet began in July 1998 and continued throughout the summer. The service model for SET appeared as a bright moment for the future of SET, but unfortunately, little interest in the service model was garnered, either, and the assets of GlobeSet were sold off in 2000.

Further interest in testing or hopes of ever rolling out SET in the U.S. had completely waned by 2000. Interest internationally, where credit card fraud very much remained a threat to the banking industry, continued to expand and several applications of SET made it into the news.

International Field Trials of SET

EasySET was one implementation of SET from the Spanish bank Banesto. EasySET was designed to answer many of the criticisms of "classical SET" by lightening the weight of consumer wallets and centralizing the complex processing of the point of sale (POS) system and the acquirer payment gateway system into a service model implementation, hosted at Banesto. This service model approach to SET takes the processing load off the merchant r-commerce systems and offers the advantages of improved transaction security and faster processing.

EasySET

Banesto's involvement with SET began in 1996 with a pilot project using the Banesto Virtual@Cash card. By mid-1997, the first Spanish SET transaction was run, and in 1999 the SET Facil, or EasySET project was launched.

The EasySET wallet supported SET transactions for Eurocard Mastercard and Visa cards issued by Banesto. The wallet was a free download to Banesto's customer using a "click-and-go" interface that enabled a 1-step download for SET cardholder certificates.

When customers used a SET-enabled credit card for payment, the EasySET POS system and payment gateway went to work at the Banesto site. Because the heavy-lifting work needed by the wallet is housed and maintained on Banesto's system, any upgrades needed to the software are made completely transparent to users. Additionally, the SafeLayer wallet supported the Electronic Commerce Markup Language (ECML) to speed up checkout processing though autofill features on merchant Web forms. Banesto also offered for free the CiberTienda shopping cart system and Virtual POS as open source downloads under the GNU public license.

In compliance with the SET specification, the EasySET systems offers the full complement of the SET messaging protocol to keep credit card information from falling into the wrong hands. It also supports the uses of SSL where SET is unavailable on cardholder registered cards. Cardholders need only download and install the SafeLayer wallet (around 500 Kbytes) and register their cards for SET enablement. Merchants download and install the Banesto Virtual POS along with the CiberTienda shopping system or within their existing e-commerce software. Multiple merchants can share the same POS software, provided that each merchant obtains and manages unique pairs of SET digital certificates needed to conduct transactions and settlement steps.

Banesto serves as the merchant CA and cardholder CA. As of mid-2002, EasySET is still in use in Spain.

Dutch Trials

I-Pay with SET was the first widespread commercial implementation of SET in the Netherlands, through Interpay Nederland B.V. in conjunction with Dutch banks offering debit or giro accounts using Maestro and Eurocard/Mastercard credit card accounts.

I-Pay offered Dutch merchants the security of SET with the additional benefit of accepting cross-border transactions from non-Dutch customers. I-Pay payments are processed within the I-Pay wallet when a customer selects either his or her Maestro account (debit) or his or her Eurocard/Mastercard (credit) as the form of payment. The I-Pay wallet prompts the user to enter a password to unlock the wallet, checks the balance on the account or the open-to-buy amount, and then challenges the user to prove legitimate ownership of the account. This step occurs in one of two ways: either a SmartCard protected by a PIN is required or a bank-supplied digital token is used to generate a one-time password once the correct secret is entered by the buyer. Two-factor authentication (what one has plus what one knows) of the buyer is sufficient proof for the banks that honor I-Pay with SET so that completed payments are irreversible, helping merchants gain confidence in debit and credit card payments via the Internet.

Merchant Web servers must run a piece of software called the Digital Till Point of Sale (POS) system to communicate with both I-Pay wallet and I-Pay payment acquirers. To accept I-Pay payments, merchants must also enter into connection agreements with Maestro for debit or giro accounts and with Eurocard/Mastercard for credit card transactions. As a merchant bank, Interpay Nederland B.V. offers both types of accounts to Dutch companies.

Struggles to Keep SET Pertinent

Suggestions for differing mechanisms to implement SET continue to crop up, including one called merchant-originated SET, or MOSET. The approach with MOSET eliminates cardholder certificates and reuses many of the traditional payment system processes with SET messages passing between the merchant POS and the acquirer, whereas the cardholder uses SSL to communicate with the merchant. Other proposed changes to SET for Version 2.0 attempted to address some of the concerns with SET Version 1.0 and add some new features, including these:

Japanese payment option (JPO) to support extended character sets,

Merchant-originated authorizations,

Online personal ID (PIN) extensions, and

CVV2/CVC2 extensions to accommodate the new Visa card fraud prevention schemes.

Other enhancements to SET that were planned included these:

Support for chip cards (smart cards);

An architecture to support debit cards with SET; and

Chip electronic commerce (CEC) to add SET messaging to the Europay, Mastercard, and Visa (EMV) specifications for chip-card payments on EMV-compatible terminals. CEC was designed as an extension to EMV (prevalent in POS systems in Europe and South America) for use on the Internet.

SET Version 2.0 never saw the light of day.

By the middle of 2002, SET had completely failed to catch on in the U.S. and it continues to languish in international markets, even as it struggles. In Spain, the Netherlands, and Finland, SET appears to have gained some traction, as dozens of merchants are SET-compliant, and the market appears to be growing.

LESSONS LEARNED AND NEW DIRECTIONS IN SECURE ONLINE PAYMENTS

The card associations eventually arrived at the realization that SET would not succeed in the U.S., and took what they learned from the experience, and went back to the drawing board. In the fall of 2001, Visa emerged with a new specification to support secure online credit card payments, called Verified by Visa, or VbV. VbV is based on the 3D-Secure payer authentication protocol, designed to authenticate cardholder identities—in real time—at the

point of purchase prior to accepting the card as payment. Meantime, Mastercard emerged in early 2002 with its own standard for Mastercard-branded products, called Secure Payments Application, or SPA.

Verified by Visa

To implement VbV, merchants are only required to install Visa-supplied software to activate a cardholder interface that challenges cardholders for passwords or asks them to insert their Visa smart cards into smart card readers and enter PINs.

Cardholders register for Verified by Visa with their participating issuer banks and can use the service when shopping at merchants that are enrolled for Verified by Visa through their merchant acquirer banks.

Verified by Visa works with traditional magnetic-strip cards using a password to identify the cardholder and also with Visa smart cards using cryptographic processing on the chip that can only be activated with the correct entry of a PIN at the time of purchase. Issuer banks must be enrolled in Verified by Visa for cardholders to use the service; otherwise the credit card payment is processed as a traditional card-not-present transaction.

The overall objectives of Verified by Visa are to improve the security of e-commerce payment transactions and to improve both cardholder and merchant confidence in Internet purchases, as well as to reduce disputes and fraudulent activity related to the use of Visa payment cards.

Components Within Verified by Visa

Verified by Visa consists of the following components that support cardholder enrollments with issuer banks and cardholder authentication to determine payment authorization:

Merchant Commerce Server: Hardware and software to support online transactions and facilitate communication between the merchant application and the merchant's acquirer bank.

Merchant Software: Software integrated into the merchant's e-commerce environment that enables merchants to participate in the Verified by Visa service.

Validation Server: Software that verifies issuer identity on digitally signed authentication responses sent to the merchant. Merchants integrate this software into their commerce server software.

Directory Server: Identifies participating Verified by Visa Issuers and cardholders and routes authentication dialog between merchants and the appropriate issuer access control server. This server is operated by Visa.

Transaction Manager Server: Stores transactions in the transaction manager database for which authentication was performed. The database is used to verify authenticated transactions and to provide additional information during the dispute process. This server is operated by Visa.

Visa Integrated Processing (VIP) Systems: Provides authorization, clearing, and settlement services through VisaNet for Visa members.

Issuer Access Control Server (IACS): Stores information about enrolled cardholder accounts and account access information in the account holder file (AHF). The server validates cardholder participation in the service and provides a digitally signed authentication response data to merchants. The IACS is operated by the issuer, processor, or Visa, on behalf of the issuer.

Issuer Enrollment Server: A server that manages cardholder enrollment by presenting a series of questions to be answered by the cardholder and verified by the issuer. The enrollment server is operated by the issuer, its processor, or Visa on behalf of the issuer.

Payer Authentication Processing

The seven steps below follow a transaction from initiation to completion using Verified by Visa:

Step 1. The Cardholder Makes a Purchase After merchandise selection through traditional online shopping steps, the cardholder proceeds to checkout. At checkout, the cardholder may complete the requested information in any variety of ways, including self-entered, an electronic wallet, merchant one-click, or other checkout capabilities. After the purchase information is entered, the cardholder selects the "buy" button. This activates the merchant plug-in to determine if the Visa card account participates in Verified by Visa.

Step 2. The Merchant Starts the Authentication Process The merchant plug-in identifies the account number and queries the Visa directory server to determine if the card account is enrolled in Verified by Visa. If the account number does not participate, the merchant plug-in returns the transaction to the merchant's commerce server and the merchant proceeds with a standard authorization request. If the account number participates in Verified by Visa, the Web site address of the issuer access control server is returned to the merchant plug-in.

Step 3. The Issuer Access Control Server Takes Control For participating cardholders, the merchant plug-in sends an authentication request to the issuer via the cardholder's browser. The issuer access control server displays a pop-up screen to the cardholder displaying information for that purchase and prompts the cardholder to enter his or her password. The cardholder enters the password and the issuer server verifies it. A cardholder is given a maximum of three attempts for password entry. If the cardholder is unable to correctly enter his/her password, the cardholder is prompted with the hint that was established during enrollment. The cardholder is given one last chance to enter the correct response. If answered correctly, the transaction continues as if the password was entered correctly. If answered incorrectly, an authentication failed response is returned to the merchant. If the cardholder has a smart Visa card, the issuer server also prompts for insertion of the chip card in the reader to initiate a dialogue with the chip. The smart Visa card generates a cryptogram that is sent to the issuer access control server along with the related transaction data used to generate the cryptogram. The server validates the cryptogram and determines if the

card authentication passes or fails. The card authentication results in information that is formatted into the response message.

After the password and/or smart Visa card is verified, the issuer access control server determines whether the cardholder authentication has passed or failed and formats an authentication response. The issuer server also sends a copy of the authentication response message to the authentication history server. All attempted authentication transaction responses (passed, failed, and not available) are transmitted and stored in the authentication history database.

Step 4. The Merchant Processes the Authorization Upon receiving the authentication response, the merchant plug-in verifies that the authorization response message is from a valid participating issuer. If it is verified and the issuer's authentication response contains a "passed" result, the cardholder is deemed "authenticated." The merchant plug-in returns the authentication response message to the merchant storefront software. If the merchant receives a "failed" authentication response from the issuer server, the merchant should request another form of payment from the shopper. Merchants are not permitted to submit failed authenticated purchases for authorization.

Step 5. The Merchant Acquirer Processes the Authorization The acquirer receives the authorization request from the merchant. The Verified by Visa data fields are mapped into existing VisaNet fields.

Step 6. VisaNet Verifies the Authentication and Processes the Authorization The VisaNet integrated payments (V.I.P.) system receives the authorization request containing the authentication data from the acquirer. These transactions are processed as standard service electronic commerce transactions.

Step 7. The Issuer Authorizes Internet Purchase The issuer's authorization center receives the request with authentication data and processes the transaction.

By mid-2002, Verified by Visa was implemented at a number of major online merchants and offered to the cardholders of Bank of America, First USA, and Bank One.

Secure Payment Application (SPA)

VbV is not without its critics. Mastercard claims that the VbV service adds processing times to transactions, takes customers off the merchant Web site, and adds complexity to integration woes, and have initially pledged not to support it. Instead, the SPA solution is Mastercard's answer to the card-not-present transaction problem. SPA relies on Mastercard's universal cardholder authentication field (UCAF) infrastructure to improve online security of payment transactions and reduce chargebacks for fraudulent transactions. SPA consists of these elements:

Issuer-provided SPA-enabled e-wallet,
SPA/UCAF value generation,
Cardholder authentication,
Merchant collection, presentation, and processing of SPA/UCAF data,
Acquirer acceptance and processing of SPA/UCAF data,
Banknet support to carry SPA/UCAF data, and
Authorization support of SPA/UCAF.

Universal Cardholder Authentication Field (UCAF)

UCAF is a 32-byte field with a variable data structure that is useful to support any number of authentication approaches to cardholder identities, including these:

SPA,
Biometrics,
Digital certificates,
Smartcards, and
Mobile and pervasive devices.

The flow for SPA processing follows:

Phase 0: Cardholders visit their credit card issuer Web sites, register their cards with SPA, establish passwords or PINs, and download and install SPA-enabled e-wallets.

Transaction Flow: Upon checkout, all traditional data are still collected (name, shipping address, billing address, etc.) whether they are filled in by the cardholder, entered via a wallet, or already stored by the merchant. The data are then posted to a Web page that the SPA-enabled wallet can access.

Once the SPA-wallet retrieves the data, it generates a payment authentication request and sends it to the issuer's wallet server.

Upon receipt of the data from the SPA-wallet, the issuer's wallet server challenges the identity of the cardholder using any method selected by the issuer (entry of password or PIN, insertion of smart card, etc.). If the challenge is met with a successful response from the cardholder, the wallet server generates a transaction-specific authentication token and sends it back to the SPA-wallet. This token is referred to as the SPA/UCAF.

The cardholder's wallet then populates the merchant's payment page with payment card details, optionally with the Mastercard card validation check value (CVC2), and with the SPA/UCAF token within a hidden field. The page is then posted back to the merchant Web server.

Once the merchant server receives the data, it formats an authorization request to the acquirer and sends along the SPA/UCAF token as a new attribute in the request.

The authorization request is then placed on Banknet and routed to the issuer bank for a response.

When the authorization request is received, the issuer bank validates that the SPA/UCAF is authentic and has not been previously used with a different transaction; it then issues an approval or declines the request based on the state of the underlying payment card. The response is then returned through the networks back to the merchant server for further processing of the sale.

SPA is intended to offer the digital equivalent of a physical cardholder signature on a record of charge, and bring

the Holy Grail of card-present transactions to the Internet. By the end of 2002, Mastercard International agreed to also support the 3D Secure Payer Authentication protocol and put an end to inter-association card acceptance problems.

Surrogate Credit Card Numbers

Still another approach that some issuer banks have adopted is called surrogate card numbers and appear in forms like Private Payments from American Express. With a surrogate card number, a shopper visits his or her issuer bank Web site and requests a "disposable" card number for one-time use at a merchant. The issuer bank keeps track of the real card number when it processes the authorization request and settlement records but conceals the number from the merchant site. Even in the event the merchant site is hacked and the tables of credit card numbers are copied, these surrogate payment card numbers would be rejected on a second authorization request and treated as fraudulent.

Although VbV, SPA, and surrogate payments may appear as evidence that the card associations cannot agree on a single secure payment system, implementing these approaches is easier than implementing SET could ever be. Vendors are offering systems and services in the marketplace to accommodate issuer banks and merchants with remotely hosted Web services applications and low-cost processing and overhead to support VbV, SPA, and surrogate card numbers. Software is available in the 2002 marketplace and is written to hide the various implementation details from the issuer banks and from the merchants to prevent the need for multiple systems to accomplish the same work. Clearly these moves are a step in the right direction.

CONCLUSION

Even as the SET specification continues to collect dust on the bookshelves of so many developers and bankers, SET's legacy is peppered with plenty of lessons to learn and mistakes to avoid. Still, SET is revolutionary, and over time, its resurrection in some form or another may materialize to finally bring an end to the intolerable state of Internet credit card fraud.

GLOSSARY

Abstract Syntax Notation One (ASN.1) A standard, flexible method that (a) describes data structures for representing, encoding, transmitting, and decoding data, (b) provides a set of formal rules for describing the structure of objects independent of machine-specific encoding techniques, (c) is a formal network-management transmission control protocol/Internet protocol (TCP/IP) language that uses human-readable notation and a compact, encoded representation of the same information used in communications protocols, and (d) is a precise, formal notation that removes ambiguities.

Acquiring bank A bank that does business with merchants who wish to accept credit cards. Merchants are given accounts to deposit the value of batch's card sales. The banks acquire batches of sales slips from any issuer bank cards and credit their value to the merchants' accounts.

Authorization A process whereby transactions are approved or declined by card issuers. Successful charge authorizations reduce the amount of available credit on a credit card but do not actually charge the customer or move money to the seller. Authorizations can be performed via telephone, POS terminal, or the Internet.

Batch settlement A process whereby accumulated credit card transactions are submitted for final settlement with a merchant's acquirer bank. Batches can be submitted for processing throughout the day or they can continue to grow until their value is sufficiently large and worthwhile to process.

Bolt-on application A "helper" or plug-in application program that extends the functionality of another program. SET is bolted onto existing merchant commerce servers and consumer Web browsers to provide POS functionality and e-wallet functionality, respectively.

Brand certificate authorities Trusted parties that serve a payment card brand (e.g., Visa, MasterCard) in performing the services needed for SET brand digital certificate management.

Branded payment cards Credit or charge cards that bear a company brand name (e.g., Visa, MasterCard, American Express).

Card associations Consist of operating banks that support franchises for particular payment card brands (e.g., Visa) and establish the by-laws that frame the uses of the franchise and the products within it.

Card issuer A bank or payment card company that issues branded cards to its customers.

Card-not-present transactions These are said to occur where the physical plastic card is not present for the merchant to see. These transactions are considered riskier than card-present transactions and typically occur with mail order/telephone order (MOTO) and Internet purchasing.

Cardholder A user (typically a consumer) of a credit or charge card issued by an issuer bank.

Certificate authorities Trusted parties who operate on the behalf of the SET Consortium (SETCo) and payment card brands to manage the distribution and currency of SET digital certificates.

Certificate revocation list (CRL) A mechanism that certificate authorities use to ensure that revoked certificates are not used in transactions. CRLs contain revoked certificate serial numbers, their date of revocation, the date the CRL was generated, its expiration date, the issuer name, and the serial number of the CA certificate used to sign the original certificate.

Certification The process of attesting to a person or entity's proof of identity and is performed prior to the issuance of a signed (notarized) certificate bearing the entity's public key.

Clearing A process of exchanging transaction details between a merchant bank (acquirer bank) and an issuer bank. Clearing posts charges to cardholder accounts and reconciles the merchant's batch of settlement records.

Closed loops Processing arrangements where a single company or bank owns both the cardholder relationship and the merchant relationship. American Express and Discover (Novus) are examples of closed-loop systems.

Credits Transactions that return money to a cardholder when goods are returned to a merchant for restocking or for defective products.

Data Encryption Standard (DES) SET's default symmetric key encryption algorithm, defined by Federal Information Processing Standard (FIPS) 46–2 and published by the National Institute of Standards and Technology (NIST).

Digital certificates Used to bind a person's identity to his orher public key and are generated by a trusted party (certificate authority).

Digital signatures Created using public–private key (PPK) cryptography and message digests (hashes of a message). Once a message digest is computed for a message, encrypted using the sender's private key, and appended to the original message, the result is called the digital signature for the message and proves that the message was sent by the claimed sender and that it was not altered en-route to the receiver.

Discount rates Paid by a merchant to a merchant bank as a privilege fee for using its credit card processing services. Fees are based on the value of each transaction and typically range from 1% up to 5%, depending on a number of factors, including charge volumes, risk models, size of the business, methods of submission, and merchant bank policies.

Electronic wallets (e-wallets) The cardholder's component for SET, which implements the protocol necessary from the cardholder end of a transaction and is used to help acquire and manage cardholder digital certificates.

Hashes A computation that reduces a large domain of possible values into a smaller range of values. Hash values and message digests are created using hashing functions. SET uses the secure hashing algorithm (SHA-1) as the default for hashing operations.

Interchanges Used to exchange information and money between the banks connected to it. The credit card interchange systems are managed by Visa and MasterCard to standardize data exchanges use across the globe.

Interchange fees Amounts charged to a acquirer bank by an issuer bank to compensate for the time the issuer bank needs to wait for payment between settlement time and actual receipt of bill payment from a customer.

Issuer banks Banks that extend credit to their customers (cardholders) through bank card accounts. These banks enter into contractual agreements (franchises) with Visa or MasterCard to issue their respective products.

Merchants Any business operation that accepts payment cards for goods or services. Merchants establish the privilege of accepting payment cards through relationships with acquiring (merchant) banks.

Merchant SET POS system A "bolt-on" application for merchant commerce servers that carries out the work necessary for online payment-card acceptance using the rules and messages defined by SET.

Open loops Contrasted to closed loops in that merchant and cardholder relationships are maintained by separate banks, but transactions with payment cards can still take place.

Out of band activities Activities that are performed outside the definition of the SET specification. For example, the exchange of order-detail information is conducted out of band to SET.

Payment gateway A front-end processor for acquirer authorization and settlement systems that translates SET messages to and from standard bank financial processing record formats.

Point-of-sale (POS) Refers to the technology (devices and systems) that carries out the work of authorizing and settling payment card charges wherever goods and services are exchanged.

Private keys The half of a key pair that are retained on the computer that generated the key pair as described by industry best practices. Private keys are used to encrypt and/or digitally sign messages that can be verified as legitimate if the associated public key is able to decrypt them.

Processing fees Charged to acquirer banks and merchants for the privilege of using the interchange network or for using merchant account services. Typically, processing fees are built into the discount rates.

Public keys The half of a key pair shared with message recipients to use in sending encrypted messages back to the private key holder. Typically available as part of an entity's certificate. In fact, authentication of entities is typically done via a message exchange between a client and an entity based on the ability to use the entity's public key to decrypt a message or digital signature with the corresponding private key held in secret by the entity.

Root certificate The topmost level in a tree of trust that is used to sign subordinate certificates. The SET root certificate is used in signing the brand certificates and is an activity performed by SETCo.

Root key authority A managing organization responsible for the generation, maintenance, and distribution of root certificates. For SET, SETCo is that managing body.

Secure hash algorithm (SHA-1) Used for hashing all data under SET. It is defined by Federal Information Processing Standards 180–1.

SETMark Visible proof of successful SET certification of vendor software, providing consumers with confidence that they are transacting using bona-fide SET.

Settlement A process that occurs when an acquiring bank exchanges financial information for funds from an issuer bank.

Third-party processors Companies that enter into contractual agreements with issuer and scquirer banks to process authorizations and settlement operations on their behalf. See charge processor.

Tree of trust Documents the hierarchy established for SET to manage the issuance, maintenance, and currency of SET digital certificates.

X.509 An industry standard to define the most widely accepted format for digital certificates, as specified by the CCITT. SET's version of X.509 digital certificates are a special "flavor" designed exclusively for payment cards.

CROSS REFERENCES

See *Digital Signatures and Electronic Signatures; Electronic Payment; Encryption; Guidelines for a Comprehensive Security System; Public Key Infrastructure (PKI); Secure Sockets Layer (SSL)*.

FURTHER READING

Cross, C. (2000). *Secure electronic transactions*. Retrieved November 2, 2000 from SANS Institute's Information Security Reading Room at http://rr.sans.org/covertchannels/SET.php

Grant, G. (1998). *Understanding SET: Visa International's official guide to secure electronic transactions*. New York: McGraw–Hill.

IBM Redbook Abstract. *Secure electronic transactions: Credit card payment on the Web in theory and practice*. Retrieved November 3, 2002 from http://publib.b.boulder.ibm.com/redbooks.nsf/redbookabstracts/sg244978.html

Loeb, L. (1998). *Secure electronic transactions : Introduction and technical reference*. Norwood, MA: Artech House.

Mastercard International. Retrieved November 2, 2002 from http://www.mastercardintl.com

Merkow, M. S., Breithaupt, J., and Wheeler, K. A. (1997). *Building SET applications for secure transactions*. New York: Wiley Computer Publishing.

RSA Security. *VISA, Mastercard and technology partners publish revised secure electronic transactions method*. Retrieved November 5, 2002 from http://www.rsasecurity.com/news/pr/960626.html

Secure Electronic Transactions, LLC (SetCo). Retrieved from http://www.setco.org

Verisign Corporation. Retrieved from http://www.verisign.com

Visa International. Retrieved from http://www.visa.com

Secure Sockets Layer (SSL)

Robert J. Boncella, *Washburn University*

E-COMMERCE AND SECURE COMMUNICATION CHANNELS

Overview

This chapter provides an overview of how the SSL protocol and its variant the TLS protocol are used to establish and operate a secure communication channel. It is assumed that the readers of this chapter are nontechnical in their academic background. As a result some space will be spent in explaining the background concepts necessary for a full understanding of SSL and TLS. If the reader requires more technical detail, Boncella (2000) is suggested.

This chapter has five major sections. First is a discussion of the need for and history of secure channels for e-commerce. Second is an overview of the internetworking concepts necessary to appreciate the details of SSL and TLS protocols. Third is a brief review of cryptographic concepts used in SSL and TLS. Fourth is a detailed exposition of SSL and TLS. And finally is a discussion of SSL and TLS protocol's status in e-commerce—its strengths and weakness, and possible alternatives.

Definition of E-commerce

E-commerce may be defined as the use of electronic or optical transmission media to carry out the exchange of goods and services. E-commerce in particular and e-business in general rely on electronic or optical communication in order to exchange information required to conduct business.

In an e-commerce transaction both the user and the provider of the service have expectations regarding the security of the transaction.

The user's expectation is that the service to be provided is legitimate, safe, and private: legitimate in the sense that the providers of the service are who they say they are; safe in the sense that the services or information being provided will not contain computer viruses or content that will allow the user's computer system to be used for malicious purposes; and finally, private in the sense that the provider of the requested information or services will not record or distribute any information the user may have sent to the provider in order to request information or services.

The server's expectation is that the requestor of the information or service is legitimate and responsible: legitimate in the sense the user has been accurately identified; responsible in that the user will not attempt to access restricted documents, crash the server, or use the server computing system as means of gaining illegal access to another computer system.

Both the server and the user have an expectation that their communications will be free from eavesdropping and reliable—meaning that their transmissions will not be modified by a third party.

The purpose of Web security for e-commerce is to meet the security expectations of users and providers. To that end, Web security is concerned with client-side security, server-side security, and secure transmission of information.

Client-side security is concerned with the techniques and practices that protect a user's privacy and the integrity of the user's computing system. The purpose of client security is to prevent malicious destruction of a user's computer systems, e.g., by a virus that might format a user's fixed disk drive, and to prevent unauthorized of use of a user's private information, e.g., use of a user's credit card number for fraudulent purposes.

Server-side security is concerned with the techniques and practices that protect the Web server software and its associated hardware from break-ins, Web site vandalism, and denial-of-service attacks. The purpose of

server-side security is to prevent modification of a Web site's contents, to prevent use of the server's hardware, software, or databases for malicious purposes, and to ensure reasonable access to a Web site's services (i.e., to avoid or minimize denial-of-service attacks).

Secure transmission is concerned with the techniques and practices that will guarantee protection from eavesdropping and intentional message modification. The purpose of these security measures is to maintain the confidentiality and integrity of user and server information as it is exchanged through the communication channel. This chapter focuses on a solution to the requirement for a secure channel.

Secure Channels

The Internet can be used for electronic communication. Those who use the Internet for this purpose, on occasion, have the need for that communication to be secure. Secure communication can be ensured by the use of a secure channel. A secure channel will provide three things for the user: authentication of those involved in the communication, confidentiality of the information exchanged in a communication, and integrity of the information exchanged in the communication.

SSL and its variant TLS are protocols that can be used to establish and use a secure communication channel between two applications exchanging information. For example, a secure channel may be required between a user's Web browser and the Web server the user has accessed. The paradigm example is the transfer of the user's credit card information to a Web site for payment of an online purchase. Another example would be an employee using the Web to send his or her check routing information to her employer for use in a direct deposit payroll request.

History of Secure Channels—SSLv1 to v3, PCT, TLS, STLP, and WTLS

Secure Sockets Layer (SSL) is a computer networking protocol that provides authentication of, confidentiality of, and integrity of information exchanged by means of a computer network.

Netscape Communications designed SSL in 1994 when it realized that users of its browser needed secure communications. SSL Version 1 was used internally by Netscape and proved unsatisfactory for use in its browsers. SSL Version 2 was developed and incorporated into Netscape Navigator versions 1.0 through 2.X. This SSLv2 had weaknesses (Stein, 1998) that required a new version of SSL. During that time—1995—Microsoft was developing PCT, Private Communications Technology, in response to the weaknesses of SSLv2. In response, Netscape developed SSL version 3, solving the weakness of SSLv2 and adding a number of features not found in PCT.

In May 1996 the Internet Engineering Task Force (IETF) authorized the Transport Layer Security (TLS) working group to standardize a SSL-type protocol. The strategy was to combine Netscape's and Microsoft's approaches to securing channels. At this time, Microsoft developed its Secure Transport Layer Protocol, which was a modification of SSLv3 and added support for UDP (datagrams) in addition to TCP support.

In 2002 the WAP Forum (wireless access protocol) adopted and adapted TLS for use in secure wireless communications with its release of WAP 2.0 Protocol Stack. This protocol provides for end-to-end security over wireless or combined wireless/wired connections (WAP Forum, 2002; Boncella, 2002).

An in-depth understanding of secure channels in general and SSL and TLS in particular requires familiarity with two sets of concepts. The first is how the client/server computing paradigm is implemented using the TCP/IP protocols. The second set of concepts deals with cryptography. In particular one needs to be familiar with the concepts of encryption, both symmetric and asymmetric (public key encryption), key sharing, message digests, and certification authorities.

The first set of concepts, clients/servers using TCP/IP, is discussed in the following section, and the cryptography concepts are reviewed following TCP/IP discussion. These cryptography concepts are discussed in detail in another chapter.

INTERNETWORKING CONCEPTS NECESSARY FOR E-COMMERCE

Clients and Servers

The World Wide Web (WWW or Web) is implemented by means of interconnection of networks of computer systems. This interconnection provides information and services to users of the Web. Computer systems in this interconnection of networks that provide services and information to users of computer systems are called Web servers. Computer systems that request services and information use software called Web browsers. The communication channel between the Web browser (client) and Web server (server) may be provided by an Internet service provider (ISP) that allows access to the communication channel for both the server and client. The communication of the client with a server follows a request/response paradigm. The client, via the communication channel, makes a request to a server and the server responds to that request via a communication channel.

The Web may be viewed as a two-way network composed of three components:

clients
servers
communication path connecting the servers and clients.

The devices that implement requests and services both are called hosts because these devices are "hosts" to the processes (computer programs) that implement the requests and services.

Communication Paths

The communication path between a server and a client can be classified in three ways:

an internet
an intranet
or an extranet.

An internet is an interconnection of networks of computers. However, the Internet (with an upper case I) refers to a specific set of interconnected computer networks that allows public access.

An intranet is a set of interconnected computer networks belonging to an organization and is accessible only by the organization's employees or members. Access to an intranet is controlled.

An extranet uses the Internet to connect private computer networks or intranets. The networks connected together may be owned by one organization or several. At some point, communication between hosts in an extranet will use a communication path that allows public access.

For a request or response message to travel through a communication path, an agreed-upon method for message creation and transmission is used. This method is referred to as a protocol. The de facto protocol of the Internet is the TCP/IP protocol. An understanding of the client/server request/response paradigm requires an overview of the TCP/IP protocol. The TCP/IP protocol can best be understood in terms of the open system interconnection (OSI) model for data communication.

The OSI Model and TCP/IP

The open system interconnection model defined by the International Standards Organization (ISO) is a seven-layer model that specifies how a message is to be constructed in order for it to be delivered through a computer network communication channel. This model is idealized. In practice, few communication protocols follow this design. Figure 1 provides a general description of each layer of the model. The sender of the message, either a request or a response message, provides input to the application layer.

The application layer processes sender input and converts it to output to be used as input for the presentation layer. The presentation layer, in turn, processes this input to provide output to the session layer, which uses that output as input, and so on, until what emerges from the physical layer is a signal that can be transmitted through the communication channel to the intended receiver of the message. The receiver's physical layer processes the signal to provide output to its data link layer, which uses that output as input and processes it to provide output to the receiver's network layer, and so on, until that message is accepted by the receiver.

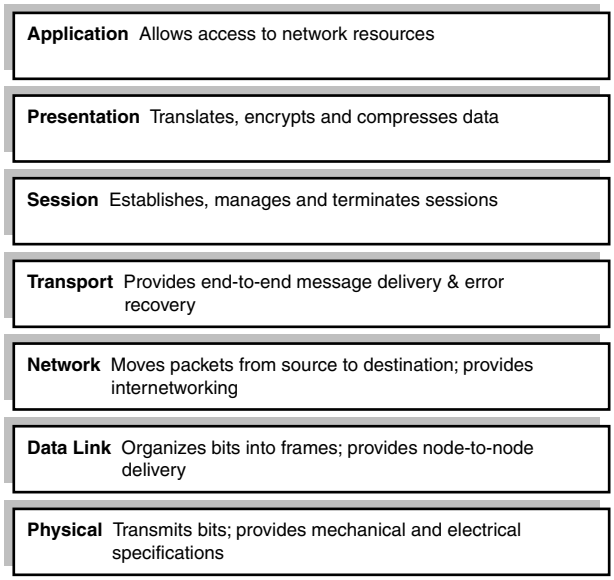


Figure 1: OSI model.

This process is depicted in Figure 2. Figure 2 also illustrates the signal (message) being relayed through the communication channel by means of intermediate nodes. An intermediate node is a host that provides a specific service whose purpose is to route a signal (message) efficiently to its intended destination.

Figure 3 depicts the TCP/IP protocol on the OSI model. (TCP/IP is an abbreviation for transmission control protocol/Internet protocol). For our purposes the TCP/IP protocol is made up of four layers. What follows is a brief overview of the TCP/IP protocol. For an introduction to the details of TCP/IP consult Forouzan (2000).

The application layer contains a number of applications that a user may use as client processes to request a service from a host. The client processes are said to run on a local host. In most cases, the requested service will be provided by a remote host. In many cases there will be a similarly named application on the remote host that will provide the service. For example, the user may open a Web browser and request HTTP (hypertext transfer protocol) service from a remote host in order to copy an HTML (hypertext markup language) formatted file into the user's Web browser. If the receiving host provides HTTP service, it will have a process running, often named HTTPD, that will provide a response to the client's request. Note that users need to specify the host by some naming method and the service they desire from that host. This is taken care of by the use of a universal resource locator (URL) (e.g., http://www.washburn.edu). The Application Layer produces a message that will be processed by the transport layer.

The client's request will pass through the local host's transport layer. The responsibility of the transport layer is to establish a connection with the process on the remote host that will provide the requested service. This client-process-to-server-process connection is implemented by means of port numbers. A port number is used to identify a process (program in execution) uniquely. Unique identification is necessary because local hosts and remote hosts may be involved in a number of simultaneous request/response transactions. The hosts' local operating systems, in concert with the TCP/IP protocol concept of port numbers, can keep track of which of several responses corresponds to the correct client process request on that local host and which request corresponds to the correct service on the remote host.

The transport layer will cut the message into units that are suitable for network transport. In addition to the port numbers, the transport layer adds information that will allow the message to be reconstructed in the receiver's transport layer. Other information is added to these units that allows flow control and error correction. The output from the transport layer is called a segment. The segment is composed of the data unit and a header containing

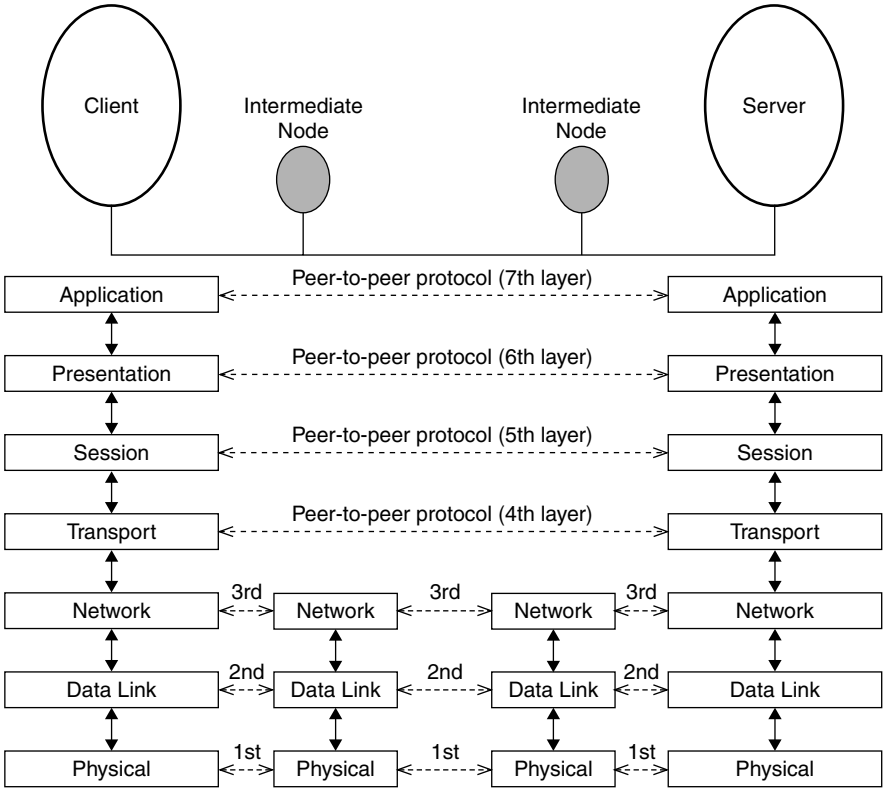
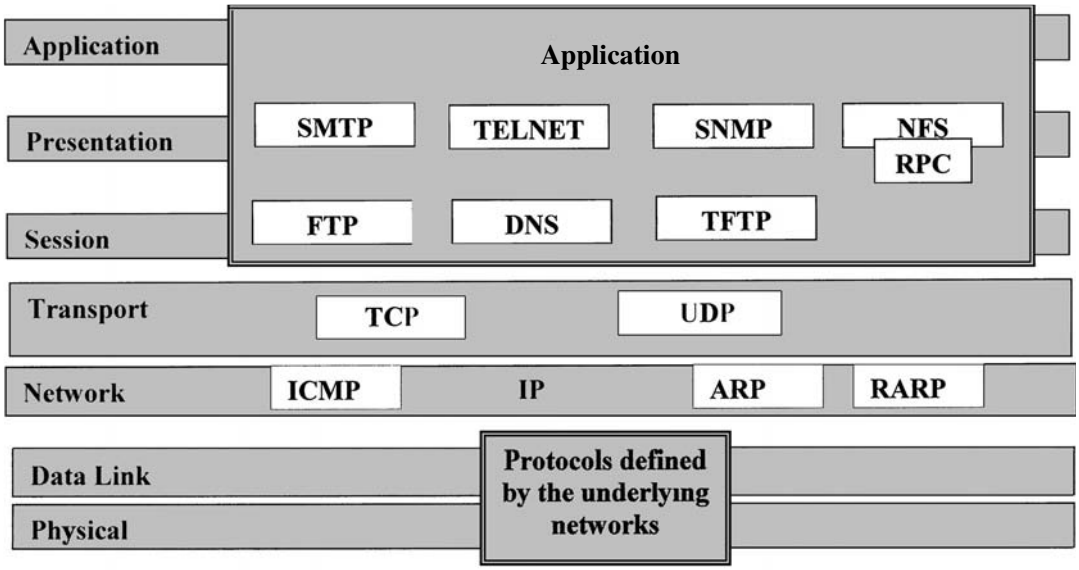


Figure 2: Messaging delivery using OSI model.



SMTP-Simple mail transfer protocol	TFTP-Trivial file transfer protocol
TELNET-Remote access program	HTTP-Hypertext transfer protocol
SNMP-Simple network management protocol	TCP-Transmission control protocol
NFS-Network file system	UDP-User datagram protocol
RPC-Remote procedure call	ICMP-Internet control message protocol
FTP-File transfer protocol	ARP-Address resolution

Figure 3: The OSI model and the TCP/IP protocol.

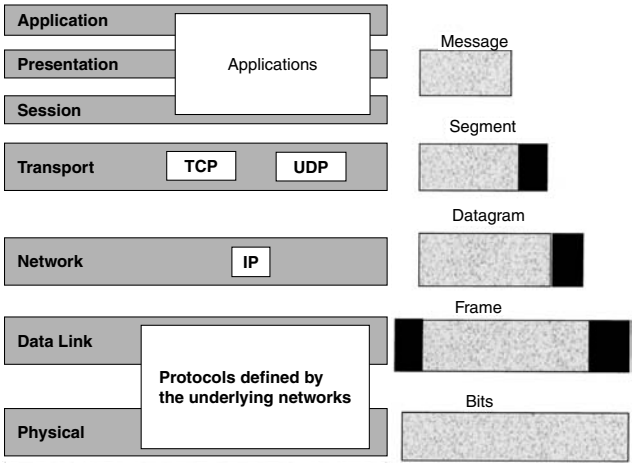


Figure 4: TCP/IP message delivery.

the information described above. Figure 4 shows this process.

The output of the transportation layer—a segment—is sent to the network or IP layer. The responsibilities of the IP layer include providing the Internet or IP address of the source (requesting) host and destination (response) host of the segment. One important part of the IP address is a specification of the network to which the host is attached. Depending on the underlying physical network, the segments may need to be fragmented into smaller data units. The information from the segment header is duplicated in each of these fragments as well as that the header information provide by the network or IP layer. The output of the IP layer is called a datagram.

The datagram is passed to the lowest layer, where the physical addresses associated with the source and destination hosts' IP addresses are added. The physical address of a host uniquely identifies the host on a network. It corresponds to a unique number of the network interface card (NIC) installed in the host. An example is the 48-bit long Ethernet address provided by the manufacturer of an Ethernet card. When the TCP/IP protocol is installed on a host, that host's physical address is associated with an IP address. The physical address allows a particular host to be independent of an IP address.

To understand Web security and e-commerce, we need to be aware of three concepts associated with the TCP/IP protocol. These are

port address
IP addresses
physical addresses.

These ideas allow the request/response message to be exchanged by the intended processes (as specified by port numbers). Those processes are running on hosts attached to the intended networks (as specified by the IP addresses) and, finally, running on the intended hosts (as specified by physical addresses). Figure 5 depicts these address assignments and the layers responsible for their assignments.

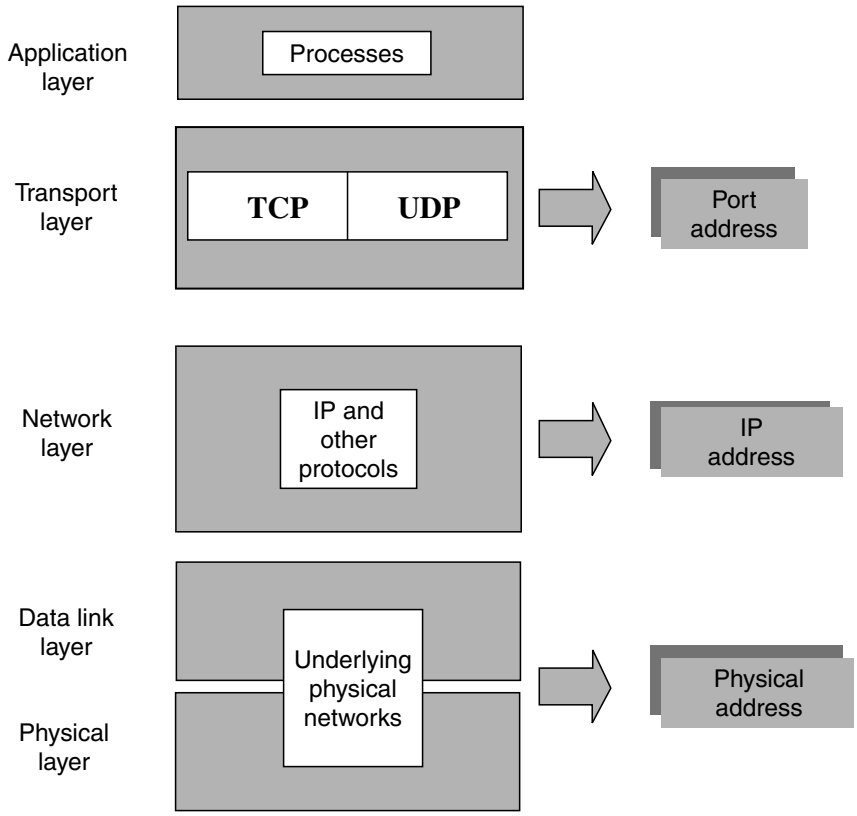


Figure 5: Address types and assignments in TCP/IP protocol.

CRYPTOGRAPHIC CONCEPTS USED IN SSL AND TLS

Encryption

Encryption is the process of converting plaintext (readable text) into ciphertext (unreadable text). Decryption is the process of converting ciphertext into plaintext. Usually this is done by means of a publicly known algorithm and a shared key. Encryption is vital in providing message confidentiality, client/server authentication, and message integrity. There are two methods of encryption: symmetric or private-key and asymmetric or public-key. Each method of encryption has its particular use. Symmetric encryption is used for encryption of the messages exchanged between a client and a server, whereas asymmetric encryption will be used to exchange the common keys used by clients and servers in their symmetric encryption process. Asymmetric encryption may also be used for the encryption of messages.

Symmetric Encryption

There are two main types of symmetric encryption: stream ciphers and block ciphers. Stream ciphers combine one byte of the key with one byte of the plaintext to create the ciphertext in a byte-after-byte process. Block ciphers process plaintext in blocks of bytes, generally 8 or 16 bytes in length, into blocks of ciphertext

RC4 is a widely used stream cipher. There are a number of block ciphers. Among them are DES, 3DES, and RC2. AES is another block cipher that is an improvement to DES. The specifics of these ciphers are discussed elsewhere in this volume.

Asymmetric Encryption

In asymmetric encryption a pair of keys, a public key and a private key, are used to carry out the encryption process. If the private key is used to create the ciphertext then only the corresponding public key can be used to decrypt that ciphertext and vice versa. Asymmetric (or public-key) encryption can be used for key sharing and digital signatures.

Key Sharing

There are two means to carry out key sharing. One is "key exchange" where one side of the message exchange pair generates a symmetric key and encrypts it with the public key of the private/public key pair of the other side. The other technique of key sharing is "key agreement." In this technique each side of the message exchange pair cooperate to generate the same key that will be used for symmetric encryption. The RSA public key algorithm can be used for the key exchange technique. The Diffie–Hellman public algorithm can be used for the key agreement technique. The details of these algorithms are discussed elsewhere in this text.

Digital Signatures

Digital signatures are used for nonrepudiation. Public-key algorithms can be used for digital signatures. RSA is a means of providing a digital signature by the sender encrypting a known pass phase with his or her private key; only the corresponding public key will decrypt the ciphertext of the pass phrase to the correct plaintext. The digital signature algorithm (DSS) is another algorithm that can be used for this purpose.

Message Digest Algorithms

Message digest algorithms are used to generate a "digest" of a message. A message digest algorithm computes a value based on the message content. The same algorithm and message content will generate the same value. If a shared secret key in included with the message before the digest is computed then when the digest is computed the result is a message authentication code (MAC). If the client and server are sharing this secret key and know each other's message digest algorithms then they can verify the integrity of the message exchange.

Two commonly used message digest algorithms are MD5, which computes a 16-byte value (128 bits), and SHA-1, which computes a 20-byte value (160 bits).

Certification Authorities

A certification authority (CA) is a trusted third party that is responsible for the distribution of the public key of a public/private key pair. The CA does this by issuing (and revoking) public key certificates. A standard for these certificates is X.509v3. This standard defines the fields contained in the certificate. This is a widely accepted standard and is used by most CAs.

SSL ARCHITECTURE

Overview

SSL is composed of four protocols. Three of the four, SSL Handshake Protocol, SSL Change Cipher Spec Protocol, and SSL Alert Protocol, are used to set up and manage secure communication channels. The remaining protocol, the SSL Record Protocol, provides the security service required by applications. The SSL lies between the application layer and the TCP layer of the TCP/IP protocols. This architecture is represented in Figure 6.

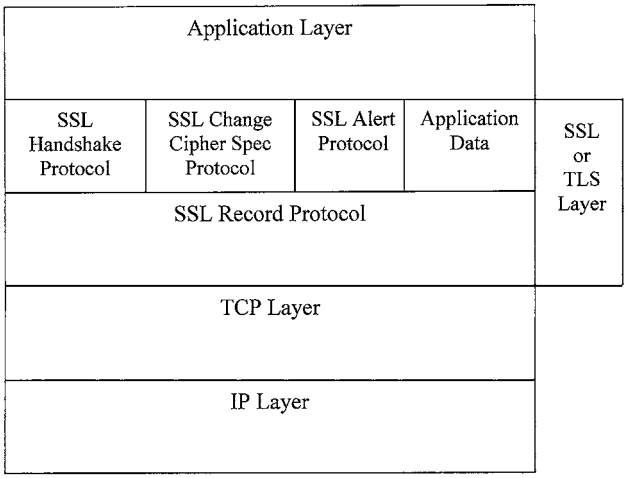


Figure 6: SSL layers within TCP/IP.

Once a secure channel has been established the SSL takes messages to be transmitted, fragments the message into manageable blocks, optionally compresses the data, applies a message authentication code (MAC), encrypts, prefixes the SSL record header, and sends the result to the TCP layer. Ultimately these data blocks are received and the data are decrypted, verified, decompressed, reassembled in the receiver's SSL layer, and then delivered to higher level clients.

The technical details of these protocols are discussed in a number of places. The primary document is the Web page http://wp.netscape.com/eng/ssl3/ssl-toc.html.

There are a number of excellent secondary sources that provide more background information as well as the specifications of the protocols. The interested reader is directed to Rescorla (2001) and Stallings (2000). The protocols used to establish a secure channel give SSL its flexibility for client/server communication.

SSL is flexible in the choice of which symmetric encryption, message digest, and authentication algorithms can be used. When an SSL client makes contact with an SSL server, they agree upon the strongest encryption methods they have in common. Also, SSL provides built-in data compression. Data compression must be done before encryption.

When an SSL connection is established, browser-to-server and server-to-browser communications are encrypted. Encryption includes

URL of requested document
Contents of the document
Contents of browser forms
Cookies sent from browser to server
Cookies sent from server to browser
Contents of HTTP header, but *not* particular browser to particular server.

In particular, socket addresses—IP address and port number—are not encrypted; however, a proxy server can be used if this type of privacy is required.

Connection Process

The connection process is shown in Figure 7. To establish an SSL connection, the client (browser) opens a connection to a server port. The browser sends a "client hello" message—Step 1. A client hello message contains the version number of SSL the browser uses, the ciphers and data compression methods it supports, and a random number to be used as input to the key generation process.

The server responds with a "server hello" message—Step 2. The server hello message contains a session ID and the chosen versions for ciphers and data compression methods the client and server have in common. The server sends its digital certificate—Step 3—which is used to authenticate the server to the client and contains the server's public key. Optionally, the server may request a client's certificate—Step 4. If requested, the client will send its certificate of authentication—Step 5. If the client has no certificate, then connection failure results. Assuming a successful connection, the client sends a

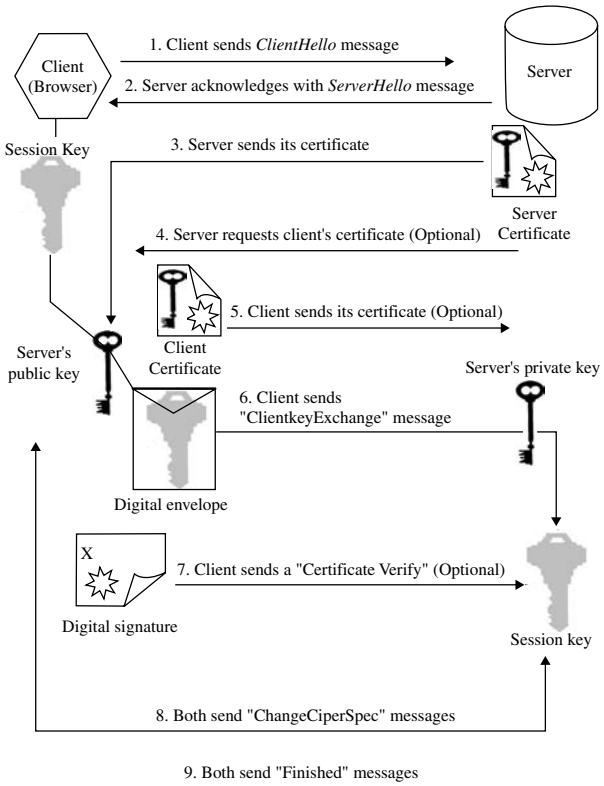


Figure 7: SSL connection process.

"ClientKeyExchange" message—Step 6. This message is a digital envelope created using the server's public key and contains the session key chosen by the client. Optionally, if client authentication is used, the client will send a certificate verify message—Step 7. The server and client send a "ChangeCipherSpec" message—Step 8—indicating they are ready to begin encrypted transmission. The client and server send finished messages to each other—Step 9. The finished messages are MACs of their entire conversation up to this point. (Note: a MAC, message authentication code, is a key-dependent one-way hash function. It has the same properties as the one-way hash functions called message digests but they have a key. Only someone with the identical key can verify the hash value derived from the message.) Accordingly, if the MACs match, then messages were exchanged without interference and, hence, the connection is legitimate.

Once the secure channel is established, application-level data can be transmitted between the client and server using the SSL Record Protocol.

Record Protocol

The SSL Record Protocol provides two of the three essential requirements for secure transmission of data: confidentiality and message integrity. Confidentiality is provided by symmetric encryption that uses the shared session key exchanged between the client and server during the handshake protocol. This handshake protocol also defines a shared secret key that can be used to create a message authentication code (MAC), which can be used

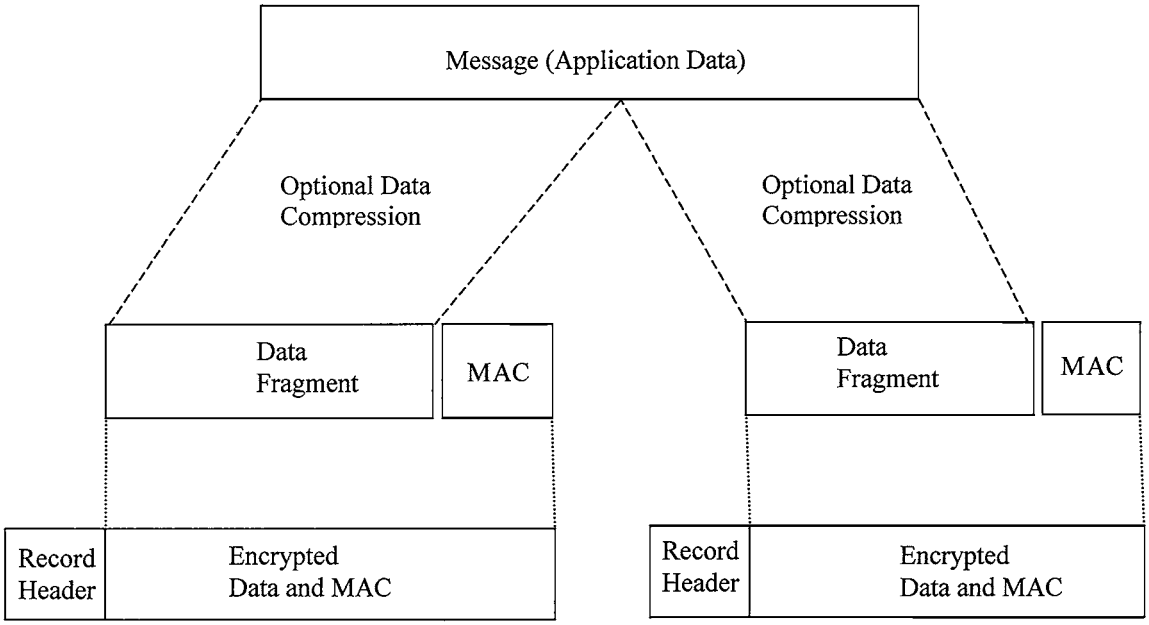


Figure 8: SSL connection process.

to ensure message integrity. The third requirement, authentication, is provided by the handshake protocol in its requirement of at least a server's certificate.

The record protocol processes a message by first breaking the message into fragments of equal fixed size, padding the last fragment as needed. The next step is optional compression of each fragment. Once the compression is completed, a MAC is computed for each fragment and appended to the fragment. The result is then encrypted using the key and algorithm agreed upon by the client and server. An SSL record header is appended. Then this segment is passed to the TCP layer for processing. The received data are processed by the receiving protocol in the reverse process: data are decrypted, verified by means of the MAC, and decompressed if necessary, the fragments are reassembled, and the result is then passed on to the destination application. This process is depicted in Figure 8.

TLS—Transport Layer Security

TLS is an IETF attempt to specify an Internet standard version for SSL. The current proposed standard for TLS is defined in RFC 2246 (2002).

The proposed TLS standard is very similar to SSLv3. The TLS record format is identical to the SSL record format. There are a few differences between SSL and TLS. Some of these are how MAC computations are carried out, how pseudorandom functions are used, including additional alert codes and client certificate types, and how certificate_verification and finished message are carried out. The details of these differences are discussed in Stallings (2000).

SSL and TLS Protocols: Details

The preceding sections provide an overview of how a secure channel is set up and used. A better understanding of this process is obtained when a detailed examination of this process is presented. It is informative to work through each step of Figure 7 and detail how the protocols work to set up the secure channel. The following is an adaptation of information that may be found in specification documents for SSL (Netscape Communications, 1996, 1998).

Handshake Protocol

Of the four protocol that make up SSL and TLS, the handshake protocol is the most critical. This protocol is responsible for setting up the connection. It uses a sequence of messages that allows the client and server to authenticate each other and agree upon encryption and MAC algorithms and their associated keys.

The format of the handshake protocol is simple and is depicted in Figure 9 below. The type field of the handshake protocol indicates one of 10 messages listed in Table 1 below. Length is the length of the message in bytes. Content is the parameters associated with the message type (cf. Table 1).

Step 1 of Figure 7 is the *ClientHello* message. Its parameters are

version The version of the SSL protocol by which the client wishes to communicate during this session. This should be the most recent version supported by the client.

random A client-generated random structure. This is a value 32 bytes long. The first four bytes are the time

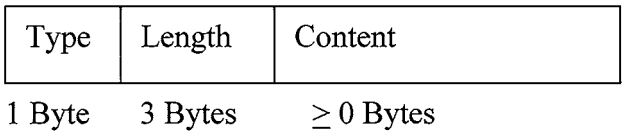


Figure 9: Handshake protocol layout.

Table 1 Handshake Protocol Messages

Message Type	Parameters
HelloRequest	Null
ClientHello	Version, random, session_id, cipher_suite, compression_method
Serverhello	Version, random, session_id, cipher_suite, compression_method
Certificate	Chain of X.509v3 certificates
ServerKeyExchange	Parameters, signatures
CertificateRequest	Type, authorities
ServerDone	Null
CertificateVerify	Signature
ClientKeyExchange	Parameters, signatures
Finished	Hash_value

of day the message was generated and the remaining 28 bytes are created using a secure random number generator. This 32-byte value will be used as one of the inputs to the key generation procedure. The time stamp (first four bytes) prevents a possible man-in-the-middle attack.

session_id The ID of a session the client wishes to use for this connection. This parameter will be *empty* if no session_id is available or the client wishes to generate new security parameters.

cipher_suites A list of the cryptographic options supported by the client, sorted descending preferences. If the session_id field is not *empty* (implying a session resumption request) this vector must include at least the cipher_suite from that session.

compression_methods A list of the compression methods supported by the client, sorted by client preference. If the session_id field is not *empty* (implying a session resumption request) this vector must include at least the compression method from that session. All implementations must support a. null compression method (i.e., no data compression is used).

After sending the ***ClientHello*** message, the client waits for a ***ServerHello*** message. Any other handshake message returned by the server except for a ***HelloRequest*** is treated as a fatal error.

Step 2 is the ***ServerHello*** message. The server processes the ***ClientHello*** message and responds with either a handshake_failure ***alert*** or a ***ServerHello*** message. The *ServerHello* message parameters are

server_version This field will contain the lower of that suggested by the client in the C*lientHello* message and the highest supported by the server.

random This structure is generated by the server and *must* be different from (and independent of) the *ClientHello* random structure.

session_id This is the identity of the session corresponding to this connection. If the *ClientHello* message session_id parameter was nonempty, the server will look in its session cache for a match. If a match is found and the server is willing to establish the new connection using the specified session state, the server will respond with the same value as was supplied by the client. This indicates a *resumed* session and dictates that the parties *must* proceed directly to the *finished* messages. Otherwise this field will contain a different value identifying the new session. The server may return an *empty* session_id to indicate that the session will not be cached and therefore cannot be resumed.

cipher_suite The single cipher suite selected by the server from the list in the ***ClientHello*** message cipher_suites parameter. For *resumed* sessions this field is the value from the state of the session being resumed.

compression_method The single compression algorithm selected by the server from the list in the *ClientHello* message compression_methods parameter. For *resumed* sessions this field is the value from the resumed session state.

Step 3 is the *Certificate* message. If the server is to be authenticated (which is generally the case), the server sends its certificate immediately following the S*erverHello* message. The certificate type must be appropriate for the selected cipher suite's key exchange algorithm, and is generally an X.509.v3 certificate. The same message type is also used for the client's response to a server's *CertificateRequest* message.

If the server has no certificate or if a key exchange technique other than RSA or fixed Diffie–Hellman is used the server will send *ServerKeyExchange* message. In this case the parameters for this message will contain the values appropriate for the key exchange technique, see (Stallings, 2000) for these details.

In **Step 4** (optional), a nonanonymous server can optionally request a certificate from the client, if appropriate for the selected cipher suite. The *CertificateRequest* massage has two parameters. These are

types A list of the types of certificates requested, sorted in order of the server's preference.

authorities A list of the distinguished names of acceptable certificate authorities.

After **Step 3** (or optional **Step 4**) the server will send a *ServerHelloDone* message to indicate that the server has sent all the handshake messages necessary for the server hello phase. After sending this message the server will wait for a client response. When the client receives the *ServerHelloDone* message the client will determine the validity of the server's certificate and the acceptability of the *ServerHello* message parameters. If the parameters and certificate are valid then the client will one or two messages.

Step 5 (optional) is the *Certificate* message. This is the first message the client can send after receiving a *ServerHelloDone* message. This message is only sent if the server requests a certificate. If no suitable certificate is available, the client should send a *NoCertificate* alert instead. This error is only a warning, however the server

may respond with a *FatalHandshakeFailure* alert if client authentication is required.

Step 6 is the *ClientKeyExchange* message. The content of the message will be based on the type of key exchange negotiated during the first phase of the handshaking process. The key exchange method is determined by the cipher suite selected and server certificate type. For example if the client and server agree upon the RSA key exchange method then the client generates a 48-byte *premaster_secret* and encrypts it with the public key from the server's certificate or uses the temporary public key from the server's *ServerKeyExchange* message.

If the server has requested a client certificate and it requires verification then the client will send a *CertificateVerify* message to provide explicit verification of its client certificate.

In **Step 8** the client sends a *ChangeCipherSpec* message that indicates the client has switched to the negotiated cipher suit. All subsequent messages will be sent using those encryption algorithms and appropriate keys. It should be noted that the *ChangeCipherSpec* message is a separate protocol and not part of the Handshake protocol. The purpose of this is to make SSL and TLS more efficient. The *ChangeCipherSpec* message consists of only one byte.

In **Step 9** the client sends the handshake message *Finish*. The message is a concatenation of two message digest values. Each value is computed using a different message digest algorithm—MD5 and SHA—on the same data. The data are the master secret (see below) and the set of handshake messages sent up to this point.

In response to these two client messages the server sends its version of the *ChangeCipherSpec* and a *Finished* message computer using that same data as the client. If this *Finished* message value differs from the *Finished* message value sent by the client then this indicates that the handshake has been modified and secure channel my not be setup. When the client receives the *finish* message from the server it does a comparison with its locally computed *finish* message value. If they match then all is well; otherwise the secure channel may not be established.

Cipher Suites and Master Secrets

There are two more concepts that need to be presented to complete this discussion. In Step 1 above the client sends a list of cipher suites to the server that the client is able to use. In Step 6 the client sends a pre_master_secret that will be used to compute the master secret. This master secret is then used to compute the key_block. This key_block is used to derive the keys that will be used with the algorithms specified in the cipher suites. The details of each of these need to presented.

Cipher Suites

A cipher suite is a list of key exchange techniques and cryptographic algorithms supported by the client and server. The **cipher_suite** parameter of the *ClientHello* message provides a set of key exchange techniques, server authentication algorithms, bulk encryption algorithms, and message digest algorithms the client can support. The client lists these sets in order of the client's preference. For example, one of the entries of this set may be

TLS_DHE_RSA_WITH_3DES_EDE_CBC_SHA

In this example the key exchange technique is DHE, where DHE denotes ephemeral Diffie–Hellman. The Diffie–Hellman parameters are signed by a DSS or RSA certificate, which has been signed by the certificate authority (CA). The signing algorithm used is specified after the DHE parameter. In this case the signing algorithm is the RSA (Rivest, Shamir, Adelman) algorithm.

The bulk encryption and message digest algorithms follow the WITH delimiter. In this the bulk encryption is performed by 3DES_EDE_CBC, where 3DES_EDE_CBD denotes 3DES encryption using the encrypt–decrypt–encrypt mode in the cipher block chaining mode, and the message digest algorithm is SHA, where SHA denotes the secure hash algorithm.

Master Secret

The master secret creation is the vital component in setting up the secure channel. The master secret is used to compute the key_block. Once the key_block computed it is partitioned into six keys that are used by the client and server in their communications. The computation of the key_block is as follows.

The *ClientKeyExchange* message provides the server with the pre_master_secret. The client and server use this 48-byte value along with the *ClientHello* random parameter value and *ServerHello* random parameter value (they both have copies of these) to create a hash value by using the MD5 and SHA algorithms in the same sequence on this common set of values. They will both compute the identical hash value. This value is the master secret that is shared (computed) by both. A similar process is used to compute the key_block but instead of using the pre_master_secret in the computation the master_secret is used. This results in a key_block that is "shared," computed independently but to the same value, by the client and server.

The size of the key_block is determined by the cipher specifications. These specifications give the number of bytes required for the bulk encryption keys (i.e., one for the client to use and one for the server to use), MAC keys, and if necessary initialization vector keys. Initialization vectors (IV) are necessary if a bulk encryption algorithm will be using the cipher block chaining mode.

This "shared" key_block is partitioned in the same sequence by the client and server. The first set of bytes are used in the client MAC secret, the next set are used for the server MAC secret, the next set are used for the client bulk encryption key, the next set for the server bulk encryption key, the next set of bytes for the client initialization vector, and finally the last set of bytes will be used as the server's initialization vector.

STATUS OF SSL

SSLv3 and TLS 1.0 and Commercial Use

SSL and TLS are primarily used to protect Web traffic that is using HTTP. In order for this to occur both the client and the server need to be SSL- and/or TLS-enabled.

Table 2 Web Servers that Support the SSL Protocol

Package	Creator	Obtain From
OpenSSL	OpenSSL Development Team	www.oepnssl.org
Apache mod_ssl (requires OpenSSL)	Apache Software Foundation	www.apache.org
Microsoft IIS	Microsoft Corporation	Bundled with WINNT, WIN2000 and WINXP
Netscape Enterprise and Suitspot	Netscape Communications	www.netscape.com
Covalent SSL (SSL Acclerator)	Covalent Technologies, Inc.	www.covalent.net
Apache Stronghold (commercial Apache)	C2Net	www.c2.net

The Web browsers Netscape Navigator and Microsoft Internet Explorer support SSL and TLS. These browsers allow the user to configure how SSL and/or TLS will be used. In Netscape Navigator 6.0 the user may consult the Security Preferences panel and open the SSL option under the Privacy and Security selection. In Internet Explorer the user may consult the Security entry in the Advanced Tab on the Internet Options selection in the drop down menu item for Tools. An interesting option in both browsers is the choice of whether or not to save the downloaded page to the local cache. The downloaded page is no longer encrypted and if it is saved to local storage it will be in plain text. If the local machine is compromised or stolen (e.g., a laptop) that document is now readable by all.

When a secure channel has been established these browsers will inform the user by means of a small padlock icon at the bottom of the browser. This indicates the page was downloaded using SSL or TLS. The URL of the web page indicates if SSL is required on the part of the web browser. A URL that begins with HTTPS indicates that SSL should be used by the browser.

A number of Web servers support SSL and TSL. A sample of such programs is displayed in Table 2.

The details of what is required to install and set up an SSL/TLS web server can be found in a number of places. For a detailed overview the reader is directed to Garfinkel & Spafford (2002) and Stein (1998). For a technical discussion of what is required the reader should consult Rescorla (2001).

Advantages and Disadvantages of and Alternatives to SSL/TLS

SSL and TLS provide server authentication, encryption of messages, and message integrity. Their design has several advantages, disadvantages, and alternatives.

Advantages

An important advantage of both SSL and TLS is they provide a generic solution to establishing and using a secure channel. This solution lies between the Application layer and TCP layer of the TCP/IP protocol suit. This implies that any protocol that can be carried over TCP (e.g., ftp, nntp) can be guaranteed security using SSL or TLS.

Another advantage is that SSL and TLS's design is publicly available. Because of this a large number of SSL and TLS implementations are available both as freeware and as commercial products. Further, these implementations are designed as APIs that are similar to networking APIs. In a C/C++-based implementation the SSL APIs emulate Berkeley sockets and in Java they emulate they Java socket class. As a result it is a simple matter to convert a nonsecure application into a secure application using SSL or TLS.

Disadvantages

In e-commerce the application of SSL and TLS has several disadvantages. Both protocols are able to solve the problem of transmitting a credit card number securely, but they are not designed to help with other aspects of that type of transaction. In particular, they are not designed to verify the credit card number, communicate and request authorization for the transaction from the consumer's bank, and ultimately process the transaction. In addition, they are not designed to carry out additional credit card services (e.g., refunds, back order processing, debit card transactions).

An additional disadvantage of SSL/TLS is security of a credit card information on the server. In particular, if the credit card number is cached on the server it will be stored in plaintext. If the server was compromised then that number would become available in plaintext.

Finally, SSL/TLS is not a global solution. In the U.S., systems that use strong encryption cannot be exported.

Alternatives to SSL/TLS

In the area of e-commerce an alternative to SSL which does not have the disadvantages cited above is SET (secure electronic transaction). SET is a cryptographic protocol developed by Visa, Mastercard, Netscape, and Microsoft. It is used for credit card transactions on the Web. It provides

Authentication: all parties to a transaction are identified;

Confidentiality: a transaction is encrypted to foil eavesdroppers;

Message integrity: it is not possible to alter an account number or transaction amount; and

Linkage: attachments can only be read by a third party if necessary.

In addition, the SET protocol supports all features of a credit card system: cardholder registration, merchant registration, purchase requests, payment authorizations, funds transfer (payment capture), chargebacks (refunds), credits, credit reversals, and debit card transactions. Further, SET can manage real-time and batch transactions and installment payments. In addition, because SET is used for financial transactions only, it can be exported and hence can be a global solution for e-commerce. The details of SET are discussed in another chapter.

In the area of providing a secure channel for messages there are alternatives to SSL/TLS.

One is IPSec (IP Security), which is a set of open standards designed by IETF and specified in RFC 2401 (2002). IPSec provides for end-to-end encryption and authentication at the IP layer. IPSec is supported in Ipv4 and mandatory in Ipv6.

Another alternative to SSL/TLS is SSH (secure shell). SSH is an application and protocol suite that allows a secure connection to be established between two computers that are using a public network. The SSH protocol architecture has three components:

Transport Layer Protocol, which provides server authentication, confidentiality, and data integrity

Authentication Protocol, which provides user authentication

Connection Protocol, which provide multiple data channels in a single encrypted tunnel.

These protocols run on top of the TCP layer in the TCP/IP protocol suite. This is similar to SSL and TLS.

GLOSSARY

Asymmetric encryption A cryptographic algorithm that uses separate but related keys for encryption and decryption. If one key of the pair is used for encryption then the other key of the pair must be used for decryption. This is sometime referred to as a public-key algorithm.

Authentication The process of verifying that a particular client or server is who it claims to be.

Block cipher A cipher that encrypts blocks of data of a fixed size.

Certificate, public key A specified formatted block of data that contains the name of the owner of a public key as well as the public key. In addition, the certificate contains the digital signature of a CA. This digital signature authenticates the CA.

Certification authority (CA) A trusted entity that signs public key certificates.

Ciphertext The result of encrypting plaintext.

Confidentiality A condition in which information exchanged between a client and server is disclosed only to those intended to receive it.

Data encryption standard (DES) A widely commercially used block cipher.

Diffie–Hellman (DH) An asymmetric algorithm that generates a secret shared between a client and server on the basis of some shared, public and randomly generated data.

Digital signature A data value computed using a public key algorithm. A data block is encrypted with the sender's private key. This ciphertext is not confidential but the message cannot be altered without using the sender's private key.

Digital signature standard (DSS) A digital signature algorithm developed by the National Security Agency (NSA) and endorsed by the National Institute of Standards and Technology.

Hash function A function that maps a variable-length message into a value of a specified bit length. This value is the hash code. There is no known method that will produce the original message using the hash value of the message. There is no known way of creating two different messages that hash to the same value.

Integrity Being able to ensure that data are transmitted from source to destination without unauthorized modification.

Internet protocol A protocol that allows packets of data to be sent between hosts in a network or hosts in connected networks.

Message digest #5 (MD5) A one-way hash algorithm.

Nonrepudiation Being able to assure the receiver that the sender of a message did indeed send that message even if the sender denies sending the message.

Rivest cipher #2 (RC2) A block cipher sold by RSA data security. This is a 40-bit key cipher.

Rivest cipher #4 (RC4) A stream cipher used in commercial products

Rivest, Shamir, Adelman (RSA) An asymmetric cipher (public-key cipher) that can encrypt/decrypt. It is also used in creating digital signatures.

Secret key A cryptographic key that is used with a symmetric algorithm.

Session key A secret key that is used for a limited period of time. This time period covers the length of time there is communication between a client and a server.

Symmetric algorithm A cipher that requires one shared key for both encryption and decryption. This shared key a is secret key and the strength of the ciphertext depends on keeping the shared key secret.

Transmission control protocol (TCP) The Internet protocol that provides reliable communication between client and a server.

Triple DES (3DES) A cipher that uses DES three times with either two or three different DES keys.

X.509 A public-key certificate.

CROSS REFERENCES

See *Authentication; Client/Server Computing; Digital Signatures and Electronic Signatures; Electronic Payment; Encryption; Guidelines for a Comprehensive Security System; Internet Security Standards; Public Key Infrastructure (PKI); Secure Electronic Transmissions (SET); TCP/IP Suite.*

REFERENCES

Boncella, R. J. (2000). Web security for e-commerce. *Communications of the AIS, 4*, Article 10. Retrieved October 1, 2002, from http://cais.isworld.org/

Boncella, R. J. (2002). Wireless Security: An Overview. *Communications of the AIS, 9*, Article 15. Retrieved March 5, 2003, from http://cais.isworld.org/

Forouzan, B. A. (2000). *TCP/IP protocol suite*. Boston, MA: McGraw–Hill.

Garfinkel, S., and Spafford, G. (2001). *Web security, privacy & commerce* (2nd ed.). Cambridge, MA: O'Reilly and Associates.

Netscape Communications (1996). *SSL 3.0 Specification*. Retrieved October 1, 2002, from http://wp.netscape.com/eng/ssl3/ssl-toc.html

Netscape Communications (1998). Introduction to SSL. Retrieved October 1, 2002, from http://developer.netscape.com/docs/manuals/security/sslin/contents.htm

Rescorla, Eric (2001). *SSL and TLS: Designing and building secure systems*. Boston, MA. Addison–Wesley.

RFC 2246 (2002). *The TLS protocol version 1.0*. Retrieved October 1, 2002 from www.ietf.org/rfc/rfc2246.txt

RFC 2401 (2002). *Security architecture for the Internet protocol*. Retrieved October 1, 2002 from http://www.ietf.org/rfc/rfc2401.txt

Stallings, William. (2000). *Network security essentials: Applications and standards*. Upper Saddle River, NJ: Prentice–Hall.

Stein, Lincoln, D. (1998). *Web security: A step-by-step reference guide,* Reading, MA: Addison–Wesley.

WAP Forum (2002). *Wireless application protocol WAP 2.0,* WAP Forum Technical White Paper. Retrieved October 1, 2002, from http://www.wapforum.org/what/WAPWhite_Paper1.pdf

FURTHER READING

Gast, M. (2002). *802.11 Wireless networks: The definitive guide*. Cambridge, MA: O'Reilly and Associates.

Netscape Communications (1999). "How SSLWorks." Retrieved October 1, 2002 from http://developer.netscape.com/tech/security/ssl/howitworks.html

Schneier, B. (1996). *Applied cryptography* (2nd ed.). New York: Wiley.

Schneier, B. (2000). *Secrets and lies: Digital security in a networked world*. New York, NY: Wiley.

Smith, R. E. (1997). *Internet cryptography*. Reading, MA: Addison–Wesley.

Stallings, W. (1999). *Cryptography and network security: Principles and practice* (2nd ed.). Upper Saddle River, NJ: Prentice–Hall.

Thomas, S. (2001). *SSL and TLS essentials*. New York: Wiley.

Viega, J., Messier, M. Chandra, and Pravir (2000). *Network security with OpenSSL*. Cambridge, MA: O'Reilly and Associates.

Securities Trading on the Internet

Marcia H. Flicker, *Fordham University*

E-FINANCE AND SECURITIES TRADING

> I don't know how the first spider in the early days of the world happened to think up this fancy idea of spinning a web, but she did, and it was clever of her, too. . . . It's not a bad pitch, on the whole. (*Charlotte's Web* [White, 1980], pp. 39–40)

Participants and observers in Wall Street's online financial web have used the term "e-finance" to name a variety of digital network technology applications—primarily using the Internet—that have transformed the personal and institutional financial markets. It has been applied to the banking, insurance, and securities industries and even to processes such as risk management in corporate finance. This chapter concentrates on online security trading and online financial services, and in this chapter, "e-finance" will refer "only" to Internet-enabled activities involved in the buying and selling of stocks, bonds, financial derivatives, and mutual funds. These activities include online investment planning, management, and trading; computerized securities exchanges; online registration of new equity offerings; and the explosion of information newly available to investors—both from commercial sources and from other investors in message boards and chat rooms. Other chapters in the *Encyclopedia* discuss online banking, electronic funds transfer, and electronic payment systems. (See Figure 1.)

With the "New Economy bubble" spinning a supportive web of capital from 1995 to 2000, the field of financial securities was transformed from one that relied on person-to-person direct communication to one that exploited the potential size, speed, and collaboration of computer networks. Technology enhanced and expedited traditional investment processes and bred new capabilities that would have been unthinkable before the World Wide Web was built.

WHY E-FINANCE?

The Industry's Perspective

> I have to get my own living, I live by my wits. I have to be sharp and clever, lest I go hungry. I have to think things out, catch what I can, take what comes . . ." (*Charlotte's Web*, p. 40)

"What comes" was more than the flies and insects Charlotte caught in her web. Three factors led businesses and governments to adopt the Internet as a distribution channel for financial services. The first two were unalloyed advantages, the third a mixed blessing:

A rapidly expanding potential market of predominantly affluent Internet users

An extremely efficient supply model for distributing information digitally

Potentially risky investments in technology infrastructures and common standards.

Potential Market

The population of Internet users has grown exponentially since the United States government released constraints on commercial applications in 1991 and user-friendly Web browsers become available in 1994. Although early users were few, they formed an attractive market segment for the financial community: comparatively affluent and innovative, and concentrated in developed and technology-rich economies such as the U.S., Canada, Northern Europe, and Australia. As the 1990s passed, the online population grew more mainstream in North America and spread to inhabitants of the developing and non-English-speaking world. According to *The UCLA Internet Report 2002—"Surveying the Digital Future"* (UCLA Center for Communication Policy, 2003), 71.1% of Americans used the Internet in 2002, whereas 47.0% of those who did not go online anticipated doing so within 12 months (pp. 18, 30). The racial and educational "digital divide" in Internet access that existed throughout the 1990s has largely disappeared; an income divide remains, both within developed economies and between affluent nations and their less affluent counterparts.

For those with access to the Net, time spent online has grown as additional products and services enhanced the utility of the Web and as surfers' experience of it deepened and matured. Years of online experience have proven to be a significant predictor of online commerce in all forms, and e-finance is no exception. *The UCLA Internet Report—Year Three* found that the average Internet user spent 11.1 hours a week online in 2002. For those with 5 years or more experience of the Web, 3.9% of that time was devoted to trading stocks, whereas those with less than a year of experience spent 2.8% of their online sessions on

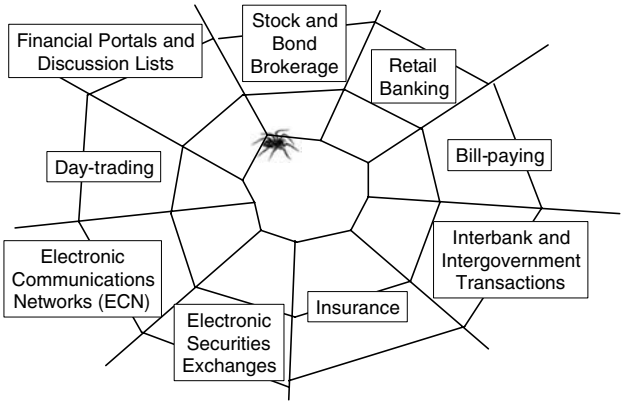


Figure 1: Wall Street's web of online financial services.

investing (p. 19). (This compares to online banking rates of 3.3% and 0.3% of online time, respectively.)

Other sources cite even greater volumes of online investing. As early as May, 1997, NetSmart announced that 42% of Internet users surveyed researched financial services online, and that 30% of them had made online investments (*Research Alert*, p. 8). The Direct Marketing Association's *Statistical Fact Book 2001* includes a Netsmart America.com study reporting that 13% of Internet users invested online in 2000 (Netsmart.America.com, 2001), and Jupiter Media Metrix forecasts 3.6 million online trades by 2006 (out of 32.5 million Internet users), up from 1.5 million in 2001 (Guglielmo, 2001). In a 2001 study, IDC estimated that there were 7 million online brokerage accounts in Europe in 2000 and forecast growth to 17 million accounts by 2004—approximately 10 million less than comparable U.S. volumes. In fact, providing online trading has become a securities industry imperative; Accenture reports that "traditional retail brokers lost $2 billion of their $54 billion in 1999 revenues to online trading companies such as E*Trade, eSchwab, and Ameritrade" (Tsien & Dumaine, 2001, p. 2).

The business-to-business financial sectors have not been left out of this revolution. ActiveMedia Research expects that "finance, insurance, and real estate" will be among the four top "Internet-based commerce leaders" in business-to-business markets by 2004, with e-commerce penetration in "transportation, trade and finance" growing from 1% in 1999 to 34% in 2004 (Karr, 2000).

Digital Distribution

Digital distribution is an extremely efficient supply model. Purely digital "products" can be sent over computer networks cheaply. It is no coincidence that the most profitable e-commerce efforts to date have not had to deal with physical goods. They were able to automate operational processes and to avoid significant warehousing, shipping, and handling expenses. Additionally, the Internet offers opportunities to automate critical procedures and to transfer many customer service activities from venders' employees to the customers themselves. In 2000, Forrester Research documented the precipitous drop in the price of information, from encyclopedias to stock prices, as the transmission medium evolved from paper and ink to bits and bytes. Online financial services were able to take full advantage of these factors. For example, after launching a revised Internet trading product in 1998—one that was low-priced but offered full access to the firm's customer services—Charles Schwab reported that it saved over $100 million annually due to "net efficiencies" (McFarlan and Tempest, 1999). (See Figure 2.)

The Investor's Perspective

> "Where do you think I'd better go?"
>
> "Anywhere you like, anywhere you like," said the goose.
>
> (*Charlotte's Web,* p. 17)

From the investor's perspective, e-finance offers opportunities unavailable in the pre-Web world. It lets individual

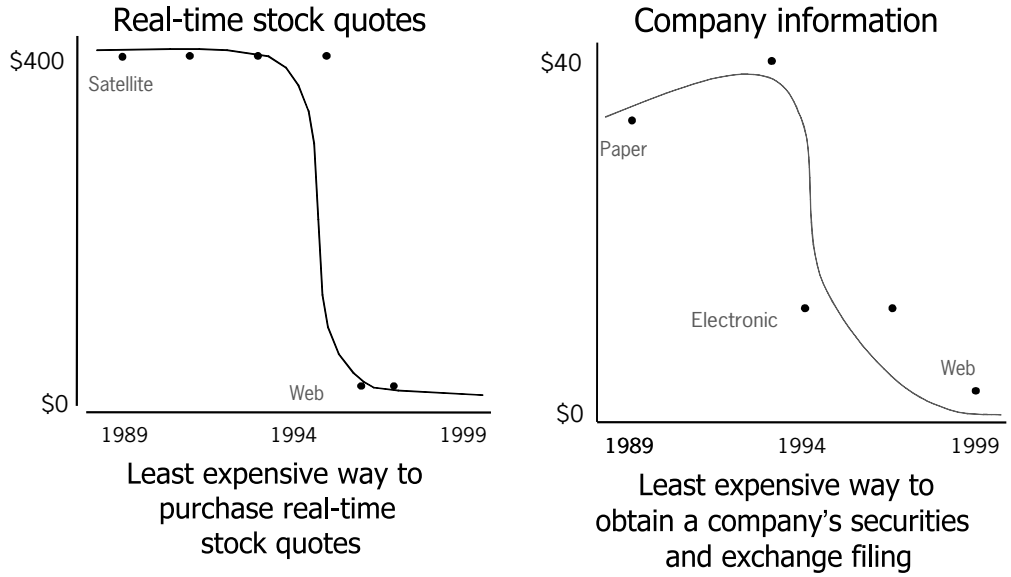


Figure 2: The cost of distributing "digital" products is minimal.

investors go almost "anywhere they like." These opportunities include:

Real-time information, which facilitates greater investment agility
Information sources beyond a human broker who may be biased by commission-driven self-interest
Low-priced trading
Membership in investor "communities" developed by specialized message boards and chat rooms.

The mass media's ubiquitous attention to finance in the late 1990s added to investors' sense of belonging, and conversely, to nonparticipants' sense of missing out on a pervasive cultural phenomenon. Only three negative factors lessened the attractiveness of online investing:

The relatively impersonal nature of online trading
Potential concerns over the security of data from both external and internal piracy—better known as "hacking"
Worries over the use or misuse of sensitive personal and financial data—the critical "privacy issue" that challenges all of e-commerce.

Real-Time, Unbiased Information

Information—voluminous and timely—is the siren call of the Internet. A variety of publishers and vendors have made financial information available online that used to be inaccessible to the individual investor, from industry and company research to real-time stock prices. Of those polled by the *The UCLA Internet Report—Year Three,* 21% cited information as their reason for starting to use the Net in the first place, making it the #1 motivator reported; 90.6% of those respondents said they considered the Internet a "moderately, very or extremely" important source of information. Their trust in the veracity of online information is not unquestioning, but it is surprisingly strong: 39.9% of Internet users considered "half" of online information "reliable and accurate" and 50.6% regarded "most" online information as reliable and accurate. Merely 7.2% endorsed only a "small portion" of online information and 0.2% believed that "none" was reliable and accurate. (Note that this question referred to all information, not financial data exclusively.)

Low-Priced Trading

From the very beginning, online stockbrokers leveraged the low cost of digital distribution into low-priced service offerings. Pioneer brokers such as E*Trade and Ameritrade passed technology-driven savings along to customers and undercut the commissions of even discount "bricks and mortar" brokers such as Charles Schwab.

Community

In addition to commercial research and professional analysis, the Internet offers virtual collaboration for gathering and evaluating information. Investors are now able to share financial news, opinions, and preferences on a variety of Web sites that offer message boards and chat rooms. It has often been said that e-commerce empowers the consumer. Online investing, by "disintermediating" the traditional broker, shifts the power—and the responsibility—for investment strategy and tactics to the individual investor. The sense of community derived from bulletin boards and chat rooms provides the personal touch that is missing from this relationship. Peer-to-peer consultations—especially when not face-to-face—allow the investor both anonymity and reinforcement. The best peer-to-peer financial sites offer basic tutorials to bring novices up to speed so that they may comfortably take part in discussions. For the knowledgeable participants, online debate and commentary can point out new opportunities or risks and can fine-tune their investment choices. Furthermore, the social value of sharing information and developing communities online has been well documented as enhancing the attractiveness and "stickiness" of a Web site by building social relationships in virtual space (Hagel & Armstrong, 1997; Martin, 2002). Many have speculated that, in a climate of escalating terrorism around the world, the need for human contact increasingly will be met though distance communications rather than through physical proximity.

Security and Privacy

Other threats, however, mitigate the physical safety of online investing. Worries about security from theft or misuse of sensitive personal information have long been barriers to Internet and e-commerce adoption. Year after year, marketing research has shown that "security" and "privacy"—often undistinguished in respondents' minds—were the primary reasons given for not exploiting the Web's shopping convenience, and they remain salient even among online shoppers and investors. The *The UCLA Internet Report—Year Three* indicated that security and privacy concerns still exist among "very experienced" (more than 5 years online) and "new" (less than a year online) Internet users alike. Of very experienced users, for example, 48.2% reported that they were "very" or "extremely" concerned about the security of their credit card data—a clear parallel to other financial information—whereas 78.6% of new users expressed that high level of concern (p. 50). (It is interesting to note that overall concern about credit card security had dropped from 2001 to 2002, with 71.3% saying they were "very or extremely concerned" about the issue in the former year and 63.3% in the latter.) Moreover, 81.6% of those already purchasing on the Internet were "somewhat," "very," or "extremely concerned" about the safety of that personal information, a privacy issue. Because most people consider personal income and wealth among the most sensitive of information categories, security and privacy must remain critical issues for e-finance providers and their customers. Disturbingly, Forrester Research found that only 70.9% of online investors were "somewhat or very satisfied" with the clarity of their primary brokerage firm's privacy policy (Table 1).

HISTORY: 1992–2002

Strands of the Web

> A spider's web is stronger than it looks. Although it is made of thin, delicate strands, the web is not easily broken. (*Charlotte's Web*, p. 55)

Table 1 North American Investors' Ranking of Brokerage-Firm Features by Overall Satisfaction* and Satisfaction with Their Brokerage Firm's Features (as a% of respondents)**

*1	Good value for the services received	67.0%**
2	Quality of financial advice off-line	62.8%
3	Financial advisors' knowledge	66.4%
4	Understanding of customer's [my] personal priorities	63.0%
5	Quality and objectivity of research	62.8%
6	Accuracy of transaction execution online	83.4%
7	[Offline] fees and commissions	57.8%
8	Stability of the institution	80.6%
9	Accuracy of account information online	88.7%
10	Accuracy of transaction execution offline	80.7%
11	Innovation of new account features or types	N/A
12	Speed of transaction execution online	79.7%
13	Online fees and commissions	67.7%
14	Helpfulness of call center representatives	69.2%
15	Accuracy of the statements	84.8%
16	Quality of financial training and education materials online	54.8%
17	Knowledge of call center representative	66.0%
18	Depth of market information online	63.8%
19	Quality of financial advice online	52.6%
20	Margin rates	41.2%
21	Ease of contacting customer service online	60.5%
22	Depth of financial research available online	61.1%
23	Speed of getting through to a call center representative	70.0%
24	A clear privacy policy	70.9%
25	Depth of account information online	79.5%
26	Speed of response to customer service requests submitted online	61.7%
27	Ability to find what customer wants on the Web site	79.0%
28	Speed of the site	76.1%

Note: Based on a survey of 1,957 North American investors.
Source: Forrester Research, March 2002. eMarketer, Inc.© 2002 (http://www.eMarketer.com)
*Asked which features most contributed to overall satisfaction with primary brokerage firm.
**Asked to indicate, about the features above, which they are somewhat or very satisfied with their primary brokerage firm.

The "thin, delicate strands" that make up the web of online financial services range from retail and institutional investors—entities such as financial portals, message boards, and day traders—to organizations that see the transactions to fruition. Participants who execute the trades include online stock brokerages, securities exchanges, newly emerged electronic communications networks (ECNs), and regulatory bodies (such as the U.S. Securities and Exchange Commission) that set the markets' rules. In a relatively short time, 10 years or less, all of these participants either have been born or have transformed their operations from a system of personal contacts (often face to face) to computerized transmission and resolution.

Day Traders

Day trading is an inherently risky, extremely short-term investment activity, with investors often buying and selling stocks within minutes in order to take advantage of rapid price changes. Professional investors had sole access to this strategy before the Internet opened it up to retail investors. Some of the purely online brokerages—such as Datek.com—specialized in serving the day trading market and developed direct trading processes that spun off as ECNs such as Island, formerly a subsidiary of Datek. Day trading reached its peak popularity from 1998 to 2000, when the bull market gave traders the illusion of invincibility. With the bursting of the dot-com bubble, however, investment activity slowed across the board as investors became more cautious. Although day trading certainly exists in 2003, it is much less prevalent than in its heyday.

Financial Portals and Message Boards

According to the comScore Media Metrix online ratings service, the top five Web properties as of July 2002 were AOL–Time Warner, Microsoft, Yahoo!, Google, and Terra Lycos. Whereas a "property" is defined as all sites owned by a given corporation, each of these domains features a gateway to financial news, and all but Google include financialdata, links, and tools as well as general-interest home pages (respectively http://www.aol.com or the ISP's welcome page, http://www.msn.com, http://www.yahoo.com, http://news.google.com/news/gnbusinessleftnav.html, and http://www.lycos.com). In addition to these sites, major news organizations such as CNN and CBS, as well as software firms such as Intuit, have created their own gateways to financial content. CNN offers

Money.CNN.com, CBS runs CBS.Marketwatch.com, and Intuit offers www.Quicken.com.

Several financial Web sites were founded with community forums at their hearts. The role of these sites is to form a convenient virtual meeting place where investors can share information and opinions with others about the economy, specific industries, and particular companies. Message boards, chat rooms, and educational content constitute the backbone of these Web sites. As media vehicles, portals and message board forums have generally partnered with online brokerages and banks in order to offer a wide range of transactional services while remaining focused on their core competencies. Two of the most consistently popular investment communities have been The Motley Fool (www.MotleyFool.com) and Raging Bull (RagingBull.Lycos.com).

Raging Bull was one of those Internet start-ups that experienced skyrocketing growth during the dot-com boom. Like Michael Dell before him, Bill Martin, founding partner of Raging Bull, turned a personal interest into a multimillion-dollar company while still in college. Having been fascinated by the stock market since age 9 and with the Internet since high school, Mr. Martin discovered early financial message board forums as a summer intern at Goldman Sachs in 1995–1996.

> As an investor I spent a ton of time that summer in the message boards. I thought, "Wow!" because I remember in high school driving 25 minutes to go to my public library to look up stocks that I owned in ValueLine. . . . And of course ValueLine only updates every couple months, . . . but I can check every day [on the Internet] and it's even cooler for these little companies you're following. A guy reads in his local paper an article and he puts it online—a little news here and there and you [put together] these tidbits and [and produce a phenomenal] amount of information. That just shows you how dramatically things have changed. It truly unleashed the amount of data and information available.
>
> I started talking to my best friend from high school—"Let's start a business together." So we started messing around at the end of '97—launched a small site. In early '98 we were kinda playing around, and then along with another guy decided that the following summer we were going to go full time with this. We took $20,000 between the three of us . . . and we launched it in June of '98 (Martin, 2002 [personal interview]).

Mr. Martin never went back to college. Within a year, Raging Bull was one of the five largest finance Web sites. Its revenue rose to almost $10 million (annualized) in 18 months. In January, 2000, it attracted 3 million unique visitors and 300 million page views. CMGI@Ventures and CNET invested $22 million. The company's management eventually decided not to go public as a stand-alone firm: "Raging Bull's community was nifty and neat, but it would be better as part of something bigger that had a whole suite of services." Instead, they sold the firm to Terra Lycos in 2000 for almost $200 million, and it become the centerpiece of Lycos' financial service offerings.

An article by Tumarkin and Whitelaw (2001) studied the applicability of message board postings as predictors of stock price and volatility. Investigating the

Table 2 Comparison of Online Brokerage Firms

	Online Revenue, November 2001	Commission on Limit/ Market Equity Order	Streaming Real-Time Data
Charles Schwab & Company	$2,461,500,000	$29.95 + $3 for order handling	Quotes, Level II, News, Charts, Time & Sales
Fidelity		$30/$25 + 2 ¢ share over 1000	Quotes, Level II
E*Trade Group	$2,171,765,000	$19.95 (limit and Nasdaq orders)/$14.95 (listed market orders) + $3 for order handling	Quotes, Level II, Watch Lists, Charts
Ameritrade*	$487,300,000	$13.00/$8.00 prior to 10/19/02, $10.95 for both thereafter	None
Datek*		$9.99	Quotes, Level II, Portfolios, Charts, Last Sale, Index Quotes
FolioFN.com		$4.00 each for trades executed two times daily, $14.95 each for real-time trades without specified price	
Sharebuilder.com		$4.00 each for trades executed at start of trading on Tuesdays, $15.95 each for real-time trades	
Buyandhold.com		$6.99 each for first 2 trades a month, $9.98 thereafter	

*Ameritrade and Datek are seeking to merge, at which time Ameritrade's fee schedule will be used.

popular belief that such community activity impacted the securities markets, the authors theorized that their influence might be due to the disclosure of new information, the reflection of market sentiment, investors' susceptibility to influence by posted messages, day traders' usage of the discussions to plumb market momentum, and consciously fraudulent efforts to manipulate the market. They found that message board discussions could be associated with short-term movement of the stocks under discussion, at least for companies in the fast-moving Internet sector, where investors could be expected to be especially vigilant. The scholars analyzed 181,633 messages taken from RagingBull.com. The 10,723 unique ticker–day combinations represented 24.1% short-term opinions and 20.8% long-term opinions. "Abnormal" stock returns for the securities discussed were defined as deviations from the Philadelphia Stock Exchange (PSE) Internet Index, and short-term abnormal returns were found to be correlated with—but not necessarily caused by—high levels of message board activity.

Online Stock Brokers

With the rise of the commercial Internet and the World Wide Web, technologically oriented entrepreneurs saw the potential benefits of online trading and launched an industry that was estimated to have captured 25% of all U.S. stock trades in 1999. Working on either a "discount" or a "deep discount" model, the earliest online brokers were "pure plays"—that is, they used the Web as their only channel of distribution to retail customers. As the 1990s ended and the dot-com bubble collapsed, the benefits of consolidation, multichannel distribution, and enriched client service became evident. Table 2 lists the top brokerage houses, in terms of their *online* revenues (i.e., excluding all other revenue) as of November 2001 and trading fees and services as of 2002. Table 3 ranks the top U.S. brokerages houses in terms of the "effectiveness" of their online offerings. The rise and stumble of online brokerage services will be detailed below.

Electronic Communications Networks (ECNs) and Stock Exchanges

Instinet, the earliest ECN, was founded in 1969 to enable institutional investors to match their large blocks of stocks directly and bypass "market makers" such as the specialists on the New York Stock Exchange (NYSE) or the dealers of Nasdaq. In 1997, the SEC imposed new regulations, called order handling rules, that required exchanges to display investors' limit orders, opening up opportunities for individual retail investors to use ECNs via their brokers. Whereas the NYSE's Rule 390 (since rescinded) limited stocks listed on the "Big Board" to trading on organized exchanges, Nasdaq imposed no such requirement. Nasdaq investors and broker/dealers were free to exploit the advantages of ECNs: low transaction fees (as low as $0.00035 per share), narrower price spreads (leading to lower purchase prices and higher sales prices), quicker execution than floor-based or screen-based systems (a fraction of a second versus half a minute or more), anonymity that offers the retail buyer the same alternatives as a large institution, and—by 1999—after-hours trading. ECNs, therefore, thrived on Nasdaq and by the first quarter of 2002 processed over 50% of Nasdaq trades (see Figure 3). Of nine ECNs founded in the past 5 years, Island was the first and remains the largest; it agreed to merge with Instinet on September 20, 2002, making their combined share of Nasdaq stock trading 22%.

ECNs are not without their disadvantages, however. Early criticism focused on their role in fragmenting the market, reducing its liquidity by shrinking the pool of potential buyers or sellers to which a given order was exposed. The larger the pool, the argument went, the greater the chance of finding an interested buyer/seller and getting/paying the best price—in Charlotte's words: the larger the web, the more likely it is to catch flies. In

Table 3 Top U.S. Brokerage Firms, Ranked by Composite Rating of Online Effectiveness (CORE) Index,* 2002 Overall Index

1	E*Trade	100
2	TD Waterhouse	100
3	ShareBuilder	82
4	Fidelity	80
5	Ameritrade	72
6	Charles Schwab	72
7	Datek	61
8	Merrill Lynch	61
9	CSFBdirect	57
10	Vanguard	48
11	American Funds	45
12	Buy and Hold	42
13	Edward Jones	40
14	American Century	40
15	Putnam Investments	34
16	PRUFN.com	29
17	T. Rowe Price	27
18	Janus	26
19	Scottrade	0.0

*The Jupiter Research CORE Index is made up of individual scores relating to number of unique visitors, usage intensity (amount of time spent), usage frequency (number of visits per month) and customer loyalty or transition (the ability to migrate off-line customers online; financial institutions that achieve the highest combination of consumers' attention, unique visitors' traffic and online transition of their total customer base will attain the highest level in the CORE ranking system.
Source: Jupiter Research, March 2002. eMarketer, Inc.© 2002 (http://www.eMarketer.com)

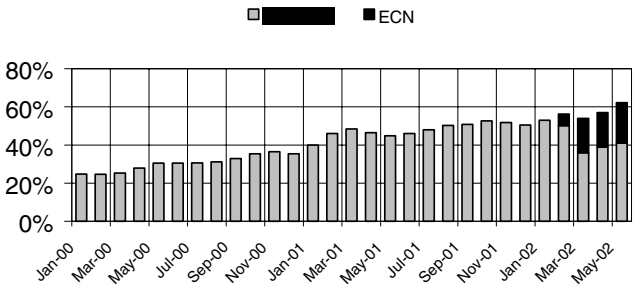


Figure 3: ECN trading volumes as percentages of NASDAQ trading volume.

order to enhance the liquidity they provided, the ECNs established mutual alliances throughout 1999 and 2000 to link their order lists and offer access to a broader market to their customers. In recent years, moreover, the field has consolidated—partially in response to increased competition from exchanges and partially due to the bear market of 2001–2003 and its lower trade volumes.

In an effort to reduce fragmentation and to defend its competitive position, Nasdaq has developed a voluntary central limit order book, known as SuperMontage, which was approved by the SEC in August 2002 and was rolled out from October 14 to December 2, 2002. Many ECNs balked at the fees Nasdaq charged as well as the competitive advantage it might have gained with the system, in which investor subscribers are notified of the best orders placed by the exchange's market-makers and any participating ECNs. Postings include both bid or asked price and the size of the offer, a piece of information that may hint at market movement. As part of its implementation, however, participants in SuperMontage give up anonymity, so users are able to infer what the big securities firms think of given stocks.

The ECNs have had a profound effect on traditional stock markets in the United States, forcing them to examine their marketing strategies and increase the value they add for customers. This has included upgrading technology significantly so that they can provide quicker order execution, enhancing the information provided to customers, and—due to competitive pressures—compressing the price spreads on securities trades. "Decimalization"—quoting prices in hundredths of a dollar instead of eighths—is one aspect of the efforts to narrow the increments among potential prices cited. In addition, exchanges that were formed as nonprofit associations have found that they cannot respond with enough flexibility to counter new competitive threats and are moving to "demutualize" and reconstitute themselves as for-profit corporations. Much of the recent revision is concentrated to the U.S.; European markets went through radical innovations that included computerization, demutualization. and collaboration in the 1980s in preparation for the economic unification that culminated with the adoption of a common currency (the euro).

Regulatory Bodies

Governments played a vital role in the growth of e-finance; they established the rules by which participants spun the web and defined the kinds of strands that would be allowed. The U.S. government was an early participant in applying technology to the securities industry by creating the initial EDGAR (Electronic Data, Gathering, Analysis and Retrieval System) registry in 1984, allowing firms to submit financial disclosure documents on computer disks. EDGAR was taken online in 1995, making detailed financial documents readily available on the Web. Moreover, the SEC's order handling rules of 1997 laid the foundation for the growth of ECNs, and later regulations opened the door for ECNs to apply for exchange status, established registration requirements for securities traded online (that is, how non-U.S. firms can qualify their Web-based offering to be exempted from registration with the SEC), and developed procedures that allowed companies to register and sell stock offerings online while bypassing underwriters (and their costs).

How the Web Was Spun

The First Strands: Discount Brokers and "Pure-Plays"

> "Well," said Mr. Zuckerman, "it seems to me you're a little off. It seems to me we have no ordinary *spider*." (*Charlotte's Web*, 80)

Early entries into the field of online stock brokerage were the discount brokerages and deep-discount brokerages that emerged from industry deregulation in the 1970s. Charles Schwab launched its first computer-based product in 1985, enabling customers to dial directly into Schwab's computer system via PC modem. E.Schwab, which was launched in 1995, was very similar to this service, still employing a proprietary telephone line to access the Schwab computer system.

Ameritrade, a pioneer in brick and mortar deep discount brokerage, was the first firm to automate consumers' trading in 1988 when it offered a touch-tone phone interface—Schwab followed in 1989—and a firm that Ameritrade later acquired (K. Aufhauser & Company) was the first to offer true Internet trading in 1994.

The first "pure-play" online brokerage—employing only the Internet for consumer trading—was E*Trade. The firm became a retail brokerage when it redirected its services from back-office online processing for discount brokers (begun in 1992) to direct-to-consumer marketing under its own brand. By 1995, commissions on consumer trades made up over 80% of E*Trade's revenue. Its long-term goal was to "become America's dominant deep-discount brokerage firm by fully automating the front and back-office trade processing function and maintaining its position as the low-cost provider" (Lal, 1996, p. 2). From 1995 to 1996, E*Trade gradually but steadily dropped its per-trade commission from $24.95 to $14.95 by exploiting its technological efficiencies. In January, 1996, it invested heavily in advertising to launch a redesigned Web site, gain brand awareness, and attract customers by positioning itself as a market innovator and technology leader with a cut-rate price. The next month, the company's advertising message evolved to differentiate itself from other deep-discounters by stressing newly added products and services: 24-hour access, free quotes, online portfolio management, free checking, and margin and I.R.A. accounts. As a result of this aggressive promotion, E*Trade was able to position itself among investors as the leading Internet broker.

In response to incursions by E*Trade and its ilk on its market share, Charles Schwab enhanced its still-limited e.Schwab service and reduced its commission to $29.95. It also increased the commission discount for its top-tier product from 10 to 20% off full-service retail. Customers and prospective customers responded positively, but as 1997 advanced, the price war among E*Trade, Ameritrade, and other deep-discounters escalated with no floor price in sight. (By 2000, some firms even experimented with free trading services.) Discussing the 2002 move by full-service brokerage houses to reject "small" clients with "only" $300,000–$400,000 to invest,

Cramer (2002) notes the changed "economics of the business":

> In the old days, as your broker, I could execute buy and sell orders for you and charge you a rate per share that could amount to as much as 25 cents on small dollar shares and as much as $1 or even $2 per share on larger amounts. If I courted you on, say, Kimberly–Clark and provided you with research and guidance about why I thought it was an appropriate time to buy the stock, and I enticed you to buy 5,000 shares at $65, I might be able to charge as much as $2,500 or $5,000 in commission. . . . But that game's dead now, slaughtered by the Net and all of those folks who charge $6 a trade!

Comparing the brokerage market to the book market, where Barnes & Noble and Borders were being cornered by an online start-up from Seattle, wits prophesied that the "brick and mortar" securities firms soon would be "Amazon'd."

Snagging the World and His Brother in the Web

> "Charlotte is fierce, brutal, scheming, bloodthirsty—everything I don't like. How can I learn to like her, even though she is pretty and, of course, clever?" (*Charlotte's Web*, 41)

As private investors achieved revolutionary access to the financial markets, their interest was reinforced by a media frenzy about the "long boom" of the 1990s and the growth of the "new economy." Market indicators and stock prices were reported and followed as enthusiastically as football scores in the final months before the Super Bowl. Even people who had never invested before began to participate in this sport.

Grass-roots participation in the equities market, combined with increased speed of execution, has been cited as causing greater volatility in stock prices and reduced holding periods during the late 1990s. In an analysis of online investor data in 2000, Roper Starch Worldwide found that the average online investor traded 12.7 times a year, with Ameritrade customers averaging 14.5 trades a year. Ameritrade itself, after examining its customer files purged of data from day-trading accounts, concluded that its customers tended to respond to short-term changes in the market.

In early 1998, Charles Schwab addressed the newly massive demand for online trading and defended its own historic positioning of value-added services at a discount by consolidating its online products into one. This product, www.CharlesSchwab.com, provided full access to Schwab research, customer service, and all communications channels for $29.95 a trade. The company also invested heavily in technology to be able to handle heavier traffic and to ensure speedy, accurate, and secure order-processing. Although the firm initially lost money and its stock price declined with the new strategy, it more than made up the difference in new customers acquired, increased trading volume among existing customers, and Internet operating efficiencies. Over the next two years, Schwab's growth, results, and market value justified the risks it took. By the end of 1999, wits were no longer talking about Barnes & Noble being "Amazon'd," but of E*Trade being "Schwabbed."

Meanwhile, traditional full-service brokers did not necessarily respond well to the challenge, fearing cannibalization of their high-fee services. Although some, such as Morgan Stanley Dean Witter, were relatively early to adopt the new distribution channel by investing in or partnering with online pure-plays and ECNs, some full-service brokers saw only the threat e-finance offered to their traditional ways of doing business. As Internet discount brokers increasingly took market share from the full-service firms, the greatest Luddite was the retail leader, Merrill Lynch. John L. Steffens, Merrill's head of retail brokerage, notoriously said in June of 1998, "The do-it-yourself model of investing, centered on Internet trading, should be regarded as a serious threat to Americans' financial lives." By the following winter, however, Merrill had spun its first tentative strands of "do-it-yourself investing" by offering a 4-month trial of free access to its global stock research on www.askmerrill.com. On June 1, 1999, it unveiled a totally redesigned strategy and announced a new multichannel vision for the firm. As Mr. Steffens himself characterized the firm's new position, "We have moved forward like a bullet train and it is our competitors that are scrambling not to get run over." Online trading had become mainstream.

Crash and Burn?

> "You lack two things needed for spinning a web . . .
>
> "You lack a set of spinnerets, and you lack know-how."
>
> (*Charlotte's Web*, pp. 58, 60)

Securities markets became increasingly shaky in the winter of 2000, and the instability culminated with a plunge in Nasdaq on April 14 that heralded the bursting of the Internet bubble. Suddenly, after the Nasdaq plunge, "do-it-yourself" investing did not appear as attractive as it had previously, especially to the relatively novice investors who had gotten into the market in the late 1990s. Issues of trust arose that undermined confidence in the quality of information provided by professionals and fellow amateurs alike. The widely quoted stock analysts of the dot-com boom were found to have had conflicts of interest after all, originating in their firms' desires to attract investment banking business from the same corporations whose potential the analysts were evaluating. "Community members" in finance forums were equally suspect: information derived from these sources could turn out to be anything from shared ignorance to outright fraud. In one notorious case, a 15-year-old New Jersey boy was caught artificially inflating the value of stocks he had purchased by posing as a knowledgeable adult and praising them in online chat rooms—a vivid demonstration of how easy it was to run a such a scam on the anonymous Web (Lewis, 2001).

Securities trading volume dropped by about 30% in 2000–2001, with the discount and deep discount

brokerages hit hardest. Newly insecure investors felt the need for reliable advice. Owing to shaky financials and an increased requirement to offer added value, the e-finance web has consolidated. There have been shakeouts, mergers, and alliances among the online discount and full-service brokerages and ECNs, providing new financial strength and access to research, recommendations, and tools that discounters had not offered in the past. Marketing strategies are evolving from a strictly low-price basis to one of convenience and personalization that leverages the nonprice strengths of the Internet.

Successful e-finance business models to date and into the future exploit multiplicity. Three business models promise a thriving potential:

Multichannel model ("clicks and mortar"): Charles Schwab successfully defended its premier industry position against online start-ups by offering its customers a variety of access points that let clients use whatever communications methods, in any combination, they chose: branch offices, telephone, e-mail, World Wide Web, and postal mail.

Multiproduct model: Financial services firms have found it far more attractive to customers, and less expensive for the firm, to offer existing clients products that span the investment, banking, and insurance industries. "Account aggregation" become the buzz phrase of 2001 as companies strove for greater "share-of-wallet" rather than more "share-of-market." E*Trade, for example, moved into the banking arena several years ago by acquiring an online bank and then established a physical footprint by buying into an ATM network.

Multiple technologies: Investors' desire for multiple touch points includes the expectation of timely information flow wherever they happen to be. Wireless reception devices—from Web-enabled cell phones to Internet-enabled PDAs (personal data assistants, hand-held computing devices)—have proliferated and become necessary accessories. Financial data are one of the services most in demand by wireless users, as seen from the list of top 10 channels in AvantGo's mobile network (Table 4).

Wall Street's web of online securities trading has been built strong but flexible. Its shape is evident, but it is equally evident that new strands are being added constantly, creating a richer and more complex net for the future. Charlotte's children may still need to struggle, but they are building an infrastructure that will last.

> Life is always a rich and steady time when you are waiting for something to happen or to hatch. (*Charlotte's Web*, p. 176)

GLOSSARY

Sources: McFarlan and Tempest (1999); Glew, Schwartz, Palumbo, Lotke, M., and Lal (1996); http://www.morganstanleyindividual.com/customerservice/dictionary/default.asp (2002); and http://www. contingencyanalysis.com/glossaryamericanoption.htm (2002).

Abnormal returns If an investment yield return on investment higher (or lower) than would be predicted by an efficient market model, it is said to have earned "abnormal" returns.

Bear market A bear market is sometimes described as a period of falling securities prices and sometimes, more specifically, as the point at which prices have fallen 20% or more from a high.

Bid and ask Bid and ask is better known as a quotation or quote. Bid is the price a market maker or broker offers to pay for a security, and ask is the price at which a market maker or dealer offers to sell. The difference between the two prices is called the bid–ask spread, or simply the spread.

Bond Bonds are debt securities issued by corporations and governments. Because most bonds pay interest on a regular basis, they are also described as fixed-income investments.

Bull market A prolonged period when stock prices as a whole are moving upward is called a bull market, although the rate at which those increases occur can vary widely from bull market to bull market. So can the length of time a bull market lasts.

Chat room This rather generic term has come to describe one of the more popular activities on the

Table 4 Top AvantGo [Wireless] Channels, Based on Units of Downloads at Advantgo.com, October 2002

	Overall Top 10	Top 10 Business/Finance
1	USATODAY.com	CNETnews.com
2	CNETnews.com	The Wall Street Journal
3	Espedia To Go	Yahoo!
4	The Wall Street Journal	CNNmoney
5	New York Times	Bloomberg
6	The Weather Channel	Business Week Online Handheld Edition
7	Yahoo!	Fool.com—Quotes and News (formerly Motley Fool)
8	CNN	FT.com
9	MSNBC.com Headlines	Zdnet to Go
10	CNN/Sports Illustrated	Economist.com Mobile Edition

Source: https://my.avantgo.com/browse/, retrieved October 24, 2002.

Internet. Using special software, Internet users can enter chat areas or "virtual spaces," where they can communicate in real time (live).

Churning If a broker buys and sells securities in an investment account at an excessive rate, it's known as churning. One indication that an account is being churned is that paymentsin commissions exceed earnings on investments. Churning is illegal but is often hard to prove.

Day trader When investors buy and sell investments within a very short time, sometimes as short as a few minutes or perhaps a few hours, they are considered day traders. The strategy is to take advantage of rapid price changes to make money quickly. In the past, professional investors did most of the day trading, but as online trading has gained popularity, many more individuals, usually referred to as electronic day traders, do it as well.

Decimalization The term decimalization denotes the move by United States securities markets to quote stock prices in hundredths (pennies) rather than eighths of a dollar.

Demutualize In an effort to become more flexible and better able to compete with ECNs and adapt to the demands of globalization, traditional stock exchanges—formed as mutual, not-for-profit associations—are switching to a corporate, for-profit structure. European exchanges, facing competition fueled by market and currency unification for two decades, were quicker to adopt this transformation than American exchanges.

Digital divide The disparity in computer and Internet access between rich and poor, ethnic minorities and majority citizens, and developed and developing countries has been called the "digital divide." It portends an increasing gap between "haves" and "have-nots," as the latter are locked out of the benefits of access to online information and services.

Discount broker Brokerages that offer securities trading at per trade commissions ($25–$35) moderately lower than traditional, full-service brokers' current fees, which were originally charged per share traded. Pioneered by the Charles Schwab Corporation in 1975, they offer independent financial products and services rather than actively managing clients' investment portfolios, and offering proprietary products and research. "Deep discount brokers" generally charge $6–$15 per trade.

Disintermediation In the early days of the commercialization of the Internet, it was widely believed that e-commerce would ultimately eliminate "middlemen" from channels of distribution by offering more desirable and more efficient direct distribution between manufacturer or service provider and end user (consumer).

Dot-com bubble The long bull market of the 1990s led to theories of a "new economy." Stock valuation for start-up, usually unprofitable, Internet firms ("dot-coms") often exceeded that of long-established and profitable "old economy" businesses in a classic investment "bubble." By the first quarter of 2000, investors' patience with red ink had worn thin and technology and Internet-sector stocks fell dramatically, most famously on April 14.

EDGAR EDGAR stands for "Electronic Data, Gathering, Analysis and Retrieval System," and was launched by the Securities & Exchange Commission (SEC) in 1984 to automate the submission and processing of financial data filings. EDGAR Online offers clients Web-based access to business, financial, and competitive information disclosed in SEC filings throughout the year by over 15,000 U.S. public companies.

Electronic Communications Network (ECN) An ECN is an alternative securities trading system that collects, displays, and executes orders electronically without a middleman (such as a specialist or market maker).

Financial portal Financial portals are Web sites that provide a single point of access to information, databases, tools, and related Web pages that help consumers manage their personal finances. Most now offer both investing and banking content.

Floor broker Floor brokers are members of a stock or commodities exchange who handle client orders that are sent to the floor of the exchange from the trading department or order room of the brokerage firms they work for.

Full service broker A full-service brokerage participates in all aspects of the investment process, from recommending investment choices to executing the transaction, measuring results, and formulating follow-up strategies. Discount brokers contend that there is an inherent conflict of interest in the full-service brokers' recommendations, as they derive revenue from trading commissions.

Individual retirement account (I.R.A.) These tax-deferred retirement accounts allow anyone who earns income from work, or is married to someone who does, to put up to $2,000 per year in an account and postpone paying tax on any earnings.

Limit order When an investor gives a broker an order to buy or sell a stock when it reaches a certain price or better, it is called a limit order. For example, if an investor places a limit order to buy a certain stock at $25 a share when its current market price is $28, the broker will not buy the stock until its share price is at $25 or lower.

Liquidity If an investment can be converted easily and quickly to cash, with little or no loss of value, it has liquidity.

Margin Buying on margin is borrowing from a broker to buy stocks. The margin is the value of the cash or securities that the buyer must deposit as collateral in a margin account. If the value of the margin account drops below the maintenance requirement, the buyer must, in most cases, add cash or securities to the account to bring its value back to the minimum.

Market maker A dealer in an electronic market, such as the Nasdaq Stock Market (Nasdaq), who is prepared to buy or sell a specific security—such as a bond or at least one round lot of a stock—at its publicly quoted price is called a market maker. Typically, there are several market makers for each security. On the floor of an exchange, such as the New York Stock Exchange (NYSE), however, the dealer who handles buying and selling a

particular stock is called a *specialist*, and there is only one specialist in each stock. Brokerage firms that maintain an inventory of a particular security to sell to their own clients, or to brokers at other firms for resale, are also called market makers.

Message board Also referred to as "discussion lists" and "bulletin boards." Web-based message boards allow users to publish questions, responses, and announcements for others to see and respond to at a later time. Unlike chat rooms, the communication is not necessarily live.

Mutual fund A mutual fund is a professionally managed investment that pools the capital of thousands of investors to trade in stocks, bonds, options, futures, currencies, or money market securities, depending on the investment objectives of the fund.

Nasdaq National Market (Nasdaq) The Nasdaq national market is part of the electronic Nasdaq stock market administered by the National Association of Securities Dealers (NASD). Stocks traded on this market must meet specific listing criteria for market capitalization and trading activity.

New York Stock Exchange (NYSE) The NYSE is the largest equity exchange in the world. Founded in 1789, it has a global market capitalization of over $15 trillion. Common and preferred stock, bonds, warrants, and rights are all traded on the NYSE, which is also known as the Big Board.

Option Buying an option gives an investor the right to buy or sell a specific investment at a specific price, called the strike price, during a preset period of time. An American-style option is an option that the holder may exercise at any time up to and including the option's expiration date. A European-style option is one that can only be exercised on its expiration date.

Over the counter (OTC) The majority of stocks in the U.S. (as well as government and municipal bonds) are traded over the counter, rather than on the floor of an organized stock exchange. That number includes more than 5,000 stocks that are listed on the Nasdaq Stock Market (Nasdaq) and are part of the National Market System (NMS), as well as stock in companies too small to meet stock market listing requirements.

Pure-play A firm is a pure-play if its only distribution channel is the Internet or the wireless Web. In the 1990s, many Internet start-ups were pure-plays.

Securities and Exchange Commission (SEC) The SEC is an independent federal agency that oversees and regulates the securities industry in the U.S. and enforces securities laws. It requires registration of all securities offered in interstate commerce and of all individuals and firms who sell those securities.

Share of market Share of market is a traditional measure of marketing success, calculated as a given company's sales divided by the sales of all competitors (including that company) in a given product market. In contrast, *share of wallet* concentrates on the individual customer. It is calculated as the percentage of an individual's purchases in a given product category that are accounted for by a given seller.

Stickiness Stickiness refers to Website content that induces visitors to spend lots of time at the site, thereby increasing their chances of responding to an advertisement or making a purchase.

Stock A stock is an investment that represents part ownership in a corporation and entitles an investor to part of that corporation's earnings and assets. Common stocks provide voting rights to shareholders but no guarantee of dividend payments. Preferred stocks provide no voting rights but guarantee a dividend payment. (Under certain circumstances and for special purposes, "restricted" nonvoting common stock may be issued by a corporation.)

Yield Yield is the rate of return on an investment, paid in dividends or interest and expressed as a percent. In the case of stocks, the yield on an investment is the dividend per share divided by the stock's price per share. With bonds, it is the interest divided by the price.

CROSS REFERENCES

See *Digital Divide; Internet Navigation (Basics, Services, and Portals)*.

REFERENCES

Cramer, J. (2002, June 17). The bottom line: Take my cash, please! *New York Magazine*. Retrieved August 24, 2002 from http://www.newyorkmetro.com/nymetro/news/bizfinance/columns/bottomline/6120/

Glew, C., Schwartz, M., Palumbo, M., Lotke, M., & Lal, R. (1996). *E*Trade Securities, Inc*. Palo Alto, CA: Stanford University. Retrieved May 17, 2002, from http://www.cnet.com

Guglielmo, C. (2001, November 12). Bottom line for financial firms: services. *Interactive Week*, 8. Retrieved February 27, 2002, from Ebhost database.

Hagel, J., III & Armstrong, A. (1997, March). *Net gain: Expanding markets through virtual communities*. Boston, MA: Harvard Business School Publishing.

Hallerman, D. (2002, May). Analyzing the rankings: Five research firms rate online brokers—eMarketer evaluates those ratings. An eMarketer analyst brief. New York: eMarketer.

Contingency Glossary. Retrieved August 26, 2002, from http://www.contingencyanalysis.com/glossaryamericanoption.htm

Dictionary of Financial Terms. Retrieved May 17, 2002, from http://www.morganstanleyindividual.com/customerservice/dictionary/default.asp

Karr, A. (2000, June). Internet-based business-to-business commerce market is poised to explode. *TeleProfessional*, 6, 24. Retrieved February 27, 2002, from Lexis–Nexis database.

Lewis, M. (2001, February 25). He wanted to get rich. He wanted to tune out his school-kid life. And neither his parents nor the S.E.C. was in a position to stop him. *The New York Times Magazine*, pp. 26+.

Martin, B. (2002). Retrieved May 15, 2002, from http://www.eFinanceInsider.com.

McFarlan, F. W. & Tempest, N. (1999). *Charles Schwab Corp. (A)*. Boston, MA: Harvard Business School Press.

Netsmart America.com (2001). Commercial online activities. In *Statistical Fact Book 2001*. New York: Direct Marketing Association.

The Internet: Bringing Wall Street to Main Street (2001, September). *Wall Street & Technology, 19,* 52–53. Retrieved March 16, 2002 from ProQuest database.

Tsien, P., & Dumaine, J. (2001). Coping with creative destruction in the securities industry: Planning for the future of financial firms and markets. In *Next Generation Investment Technology 2001*. Retrieved February 15, 2002, from www.accenture.com/xc/xd.asp?it=enWeb&xd = industries/financial/fsi_creative.xml

Tumarkin, R., & Whitelaw, R. F. (2001, May/June). News or noise? Internet postings and stock prices. *Financial Analysts Journal. 57*(3), 41–51. Retrieved February 19, 2002, from ProQuest database.

UCLA Center for Communication Policy Co (January, 2003). *The UCLA Internet Report 2002—"Surveying the Digital Future"*. Retrieved March 24, 2003, from http://www.ccp.ucla.edu

White, E. B. (1980). *Charlotte's Web*. New York: HarperTrophy. [Original work published in 1952.]

FURTHER READING

Angel, J. (2000). *Market Mechanics: An Educator's Guide to U.S. Stock Markets*. New York: The Nasdaq Stock Market University Outreach.

Colarusso, D. (2002, March 10). Day trading takes a conservative turn. *The New York Times,* p. BU 6.

Eagleson, J. (2002). Trading places: The capital markets' investment in straight through processing. In *Food for Thought: Straight Through Processing*. McLean, VA: KPMG Consulting. Retrieved May 13, 2001, from http://www.kpmg.com (now http://www.baringpoint.com)

Kirsner, S. (2000, November). 'The Internet is going to change Wall Street as we know it." *Fast Company*, (35), 204+. Retrieved March 16, 2002, from http://www.fastcompany.com/online/40/wf_miller.html

Kollock, P., & Jaycobs, R. (2001, April 13). Liquidity Myths. Reprinted from the January/February issue of *@Markets Magazine*. Retrieved May 13, 2002, from http://www.commercenet.com

Levitt, A. (1999). The changing markets. *Vital Speeches of the Day, 66*(1), 7–10. Retrieved February 19, 2002, from ProQuest database.

National Association of Securities Dealers Regulatory Site.Retrieved May 15, 2002, from http://www.nasdr.com

Rigby, D. (2000). Winning the turbulence—Strategies for success in turbulent times. *European Business Journal, 12*(2), 76–86.

Rosato, D. (2002, May 12). Investing: At some online brokers, discounts have a price. *The New York Times*, p. MB 7. Retrieved September 22, 2002, from Lexis–Nexis database.

Smith, G., & Schmitt, C. (2001, July 23). Time to real in the portals? *Business Week, 3742,* 70–71. Retrieved March 9, 2002, from ProQuest database.

Staff reports (1999, November 29). Internet is now leading source of investor information. *Investor Relations Business, 1,* 12.

Stake your claim to wealth: Technology guide with table of notable web sites (2001, Winter). *Fortune, 142*, 248–260.

Tully, S. (1999, August 2). Will the web eat Wall Street? *Fortune 140*(3) 112–118. Retrieved March 8, 2002, from ProQuest database.

Weinberg, N. (2001, October 1). After the bubble. *Forbes, 168*(8), 60–68. Retrieved March 9, 2002 from ProQuest database.

Wright, A. (2002). Technology as an enabler of the global branding of retail financial services. *Journal of International Marketing, 10*(2), 83–98. Retrieved August 23, 2002 from ProQuest database.

Software Design and Implementation in the Web Environment

Jeff Offutt, *George Mason University*

INTRODUCTION

The original Web sites used hyperlinks to connect text documents. Modern Web applications run large-scale software applications that support e-business, information distribution, entertainment, collaborative working, surveys, and numerous other activities. They run on distributed hardware platforms and heterogeneous computer systems. The software is distributed, is implemented in multiple languages and styles, incorporates reuse and third-party components, is built with cutting-edge technologies, and must interface with users, other Web sites, and databases. The software components are often distributed geographically both during development and deployment and communicate in numerous distinct and sometimes novel ways. Web applications consist of heterogeneous components including traditional and nontraditional software, interpreted scripting languages, plain HTML (hypertext markup language) files, mixtures of HTML and programs, databases, graphical images, and complex user interfaces. This heterogeneity has led to the notion of Web site engineering (Powell, 1998).

The tremendous reach of Web applications into all areas of communication and commerce makes this one of the largest and most important parts of the software industry. Yet studies (President's Information Technology Advisory Committee [PITAC] 1999; Schneider, 1999) have found that the current base of science and technology is inadequate for building systems to control critical software infrastructure. Web software development uses cutting-edge, diverse technologies, and we are just beginning to learn how to design and develop high-quality Web software, making this problem particularly severe.

FIRST-GENERATION WEB SITES

The original Web sites were static HTML files, so-called soft brochures, usually created by a single webmaster who used technologies such as HTML and simple CGI scripts to present information to visitors and occasionally obtain information from them with forms (Powell, 1998). Figure 1 illustrates this scenario. A client was a Web browser that people used to visit Web sites. The Web sites were on separate computers, the servers, which delivered HTML files to the client. HTML forms generated data that were sent back to the server to be processed by CGI programs. This is a simple execution model that supports relatively small Web sites. The software involved is by necessity small in scale; such Web sites usually cannot support much load and offer limited functionality. The software also has few provisions for security, and the TCP (transmission control protocol) and HTTP (hypertext transfer protocol) by themselves are not designed to support secure interactions.

SECOND-, THIRD-, AND FOURTH-GENERATION WEB SITES

This situation drastically changed through the late 1990s, with strong impact on and motivation from engineering principles and processes. Second-generation Web sites featured significantly more layout and presentation abilities, graphics, and more robust backend software support, including session management with cookies.

Third-generation Web sites added improved interaction, including client-side execution such as JavaScripts and Java applets. Third-generation Web sites also became fully functional software systems and provide business-to-customer and business-to-business e-business, and a large variety of services to a large variety of users.

Developers of third-generation Web sites found several problems with the software support for the increased level of uses. It was difficult to achieve the reliability needed for e-business, security became a problem, maintenance was difficult, and the software designs did not scale well. Fourth-generation Web sites currently rely on multitiered hardware and software architectures, improved software technologies such as the J2EE platform, communication

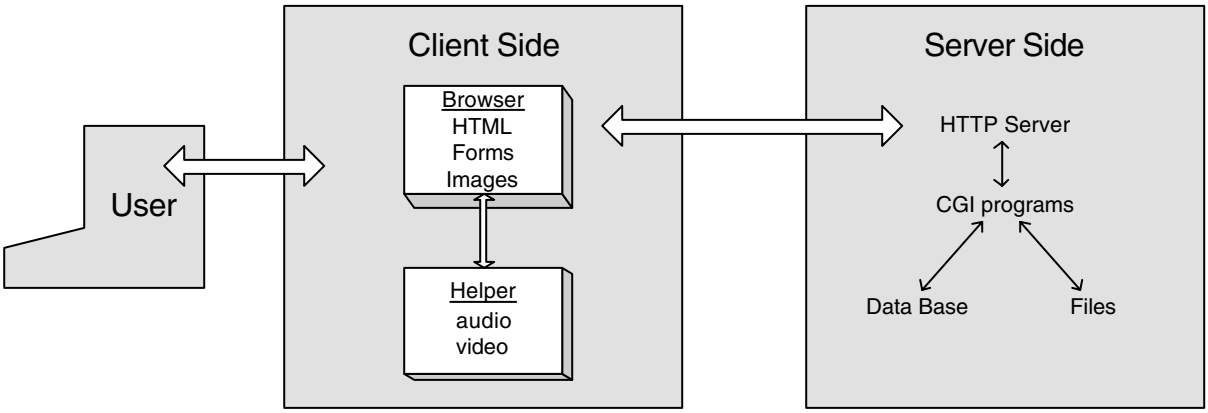


Figure 1: First-generation Web sites.

among software components with XML (extensible markup language), and a number of design architectures for large-scale Web software applications. Figure 2 illustrates a typical configuration for a fourth-generation Web application.

Most of the software has been moved to a separate computer, the application server. Large Web sites implement the application server as a collection of application servers that operate in parallel. Likewise, Web servers are often clusters of computers that work together to server requests from large numbers of users. Application servers typically interact with one or more database servers, often running a commercial database. The client–server interaction, as before, uses the Internet. The Web servers and application servers are connected by middleware, which are packages obtained from software vendors to handle functions such as communication, data translation, and process distribution. Middleware is sometimes as simple as Java Data Base Connectivity (JDBC), whereas other middleware packages are large and solve complicated problems. Likewise, the application-database servers often interact through middleware.

WEB SOFTWARE ENGINEERING QUALITY FACTORS

Although software engineering researchers, educators, and practitioners have spent years focusing on developing processes and technologies to improve software quality attributes, much of the software industry has had little motivation to improve the quality of their software. Software is often sold with relatively low-quality requirements; the combination of user expectations and market realities has been such that increasing quality usually has not increased profits. A combination of time-to-market and marketing strategies has almost always determined whether traditional software products succeed competitively. As an example, software contractors for government agencies are often paid the same regardless of the quality of the delivered software. Despite the positive impacts of the capability maturity model (Carnegie Mellon Software Engineering Institute, 2002), many contractors are still given additional resources to correct problems of their own making (Tassey, 2002). Commercial software companies (so-called shrink-wrap vendors) are usually

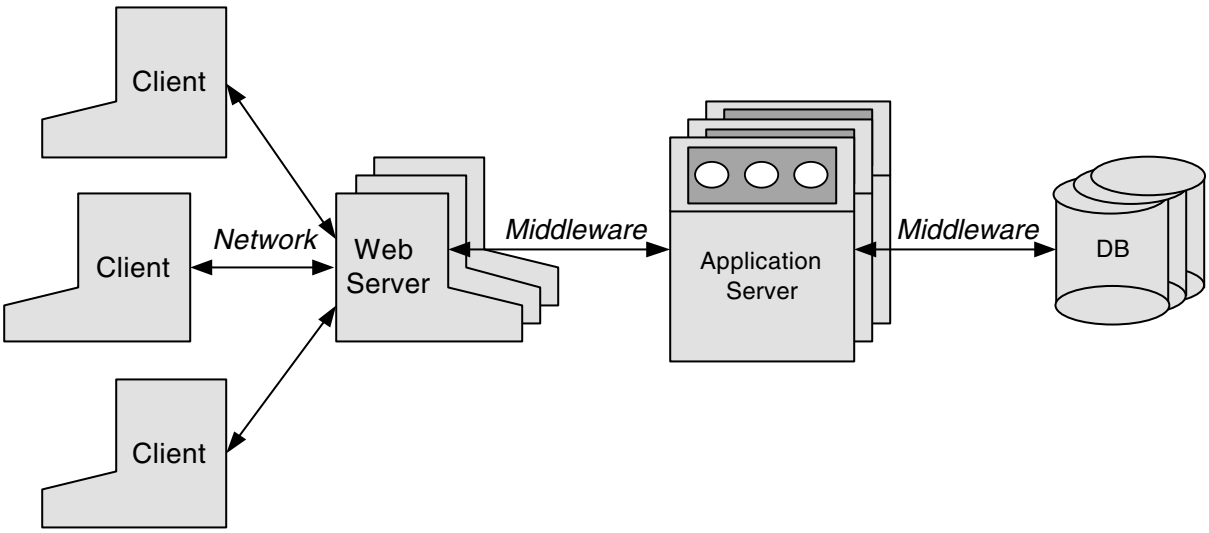


Figure 2: Multitier Web sites.

driven almost entirely by time-to-market; it is almost invariably more lucrative to deliver poor-quality products sooner than high-quality products later. It is a well-known truism that companies can often sell poor-quality first versions of software applications and then make more money by charging for upgrades that contain more bug fixes than new features. For most applications, there has traditionally been little economic motivation for producing high-quality software. In fact, there have often been economic disincentives for creating high-quality software products.

Web-based software is in a completely different situation, one more akin to critical software such as aerospace, telecommunications, and medical devices. One of the interesting challenges is that Web software has extremely high-quality requirements (Offutt 2002; Powell 1998). However, there appears to be little or no brand-name loyalty (that is, "site loyalty") for Web applications. Many companies that sell through the Web depend on customers using their sites, and most important, returning to their sites. Others offer more traditional software services that are available through the Web; these services will not be used if the quality is too low because it is relatively easy for users to switch to other services. Thus, unlike many contractors, Web site developers will only see a return on their investment if their Web sites exhibit sufficient quality (that is, if the Web sites satisfy users' needs). Unlike many software vendors, if a new company puts up a competitive site that customers perceive to have higher quality, customers will almost immediately shift their business to the new site. Thus, it is often advantageous to be "later than and better" instead of "sooner but worse." Although the idea of "sticky Web sites" has been discussed and mechanisms to encourage customers to come back have been developed (Menascé, 2000), thus far the key mechanism to bring repeat customers to Web sites is high quality. It seems likely that this will continue to be true for the foreseeable future.

In software development, a process-driver is a factor that has a strong influence on the process used to develop the software. Thus, if software is required to have high reliability, the development process must be adapted to ensure that the software works well. When I have surveyed the important quality process drivers for traditional software, developers always give a single answer that stands alone far above the rest: time-to-market. But when I recently made the same survey of Web software development managers and practitioners, they claim that time-to-market, although still important, is no longer the dominant process driver. Instead, the three most important quality criteria for success of Web applications (and thus, the underlying software) were given as

Reliability
Usability
Security

An additional four important criteria that were listed are

Availability
Scalability
Maintainability
Performance

Of course, it would be foolish to claim that this is a complete list of important or even relevant quality attributes. Quality attributes such as customer service, quality of products, price, and delivery are also important but not related to the software and thus beyond the scope of this chapter. Nevertheless, these quality attributes track closely with what is said in other books and articles (Constantine & Lockwood, 2000; Dustin, Rashka, & McDarmid, 2001; Kassem et al., 2000; Murugesan & Deshpande, 2001; Powell, 1998; Scharl, 2000). Thus, there is wide agreement that satisfying quality attributes is essential to Web software, and these seven provide a solid basis for evaluating Web software. These quality attributes are used as a basis for suggesting specific ways to engineer Web site software, using the available technologies.

Before proceeding with the technology aspects of this chapter, let's explore the reasons these quality attributes are so important. These quality factors will have a much stronger impact on the profits of Web-based companies than for most traditional software. The reasons for the first three quality requirements—reliability, usability, and security—may become obvious by analyzing some of the new uses of this software. The most obvious is that of direct selling to customers, that is, "B2C." This includes companies that sell books and other small items such as Amazon, plane tickets such as Yahoo and Expedia, and rental companies such as Netflix. Customers who buy books from a Web site expect the same quality of service that they would get from going to a bookstore at the mall, but without the overhead of actually driving to the mall. We expect to be able to find the books we want in a convenient way (usability), we expect to be able to make the purchase without difficulty (usability), we expect the correct books to arrive at our house in the specified number of days (reliability), and we expect the correct amount to be billed to our credit card (reliability).

The issue of security of Web applications is getting more important. One of the major concerns of security for e-business, of course, has to do with security of data. Customers expect their credit card and personal information to be held in confidence. Identity theft, where a criminal takes the entire credit history and assumes the name of another person, is becoming more common and can be done by taking advantage holes in Web software. The security also works in the opposite direction. Improper use of cookies has opened up holes for users. One example is that of Web software storing price information in a cookie on the client-side, which allowed customers to change the price of items they bought on the Web. As a field, we are continuing to evolve our expectations of security and our ability to support security.

The additional quality requirements are less obvious. Whereas a bookstore on the corner ("brick and mortar") might expect to have customers form the neighborhood Monday through Saturday, 8:00 a.m. to 7:00 p.m., a Web-based company can expect customers from all over the world. It might be 3:00 in the morning in Virginia, but it's the middle of the afternoon in Beijing! It might be Thanksgiving holiday in the United States, but it's just another spring day in South Africa. Thus, Web sites must have extremely high availability, not just 24/7, but 24/7/365.

Another key difference is that, unlike shrink-wrap software applications, Web-based applications do not have to be sold or distributed when updates are made. Consider maintenance updates to a commercial word processing program. Immediately after releasing one version, the company starts collecting problems and making a list of needed changes. The first change might be simple and easy, completed within a week or a few days after the version is released. That change is not made available to the customers immediately, however, but held for months or years until the company is able to release the next version. With Web software, on the other hand, that small change can be installed live immediately; moreover, customers expect it to be. These factors, together with the rapid evolution of technology, means that maintainability is crucial for Web software. Instead of an update rate of months or years, Web software must be able to support an update rate of days or even hours.

Unlike traditional businesses whose potential customer base is typically limited by physical concerns such as geography and traffic, growth in Web-based businesses has unlimited potential: There are currently hundreds of millions of users on the Web, each of whom is only a click away and therefore "in the neighborhood" of the store. This means that Web software must be highly scalable and ready to grow in terms of servers, services, and customers very quickly.

Finally, customers expect Web sites to respond quickly to their requests. Nielsen (2000) claimed that users perceive a response that occurs within 1 second to be immediate, but will lose concentration and thus interest after five seconds. After thirty seconds without a response, they will almost certainly give up. Although network speeds usually dominate response times, a bad software design can seriously disrupt performance.

These quality requirements are not new, and certain segments of the software industry have faced some of these problems in various contexts. The novel aspect is that Web software has all of these quality requirements at once. Many of the technological innovations of the past 5 years have been either in response to these requirements or to support the fundamental distributed nature of Web software.

TECHNOLOGIES FOR BUILDING WEB SITE SOFTWARE

The changes in technology for building Web software through the late 1990s and early 2000s have been continuous, fast-paced, fundamental, and dramatic in scope. These changes continue, thus this chapter can only provide a snapshot of the current technologies that are being used. Several varieties of plug-in enabling technologies are currently used to support Web software. An enabling technology is generally any mechanism that is used to make Web pages dynamic and respond to user input. Web browsers use plug-in modules to handle specific enabling technologies on the client. Web servers use server-modules to handle enabling technologies on the server.

Two common varieties of plug-ins to support server-side processing are compiled modules and scripted pages. Compiled modules are executable programs that the server uses to support server-side processing. Compiled modules currently in common use are Java servlets, Apache Modules, Microsoft's Internet server application program interface (ISAPI), and Netscape's server API (NSAPI). Scripted pages are HTML pages that also have the ability to process business logic. Scripted pages are executed server-side, not client-side (as JavaScripts are), but they are HTML pages that can access software on the server to get and process data. Scripted pages currently in common use are Java Server Pages (JSP), Macromedia's Cold Fusion, Microsoft's Active Server Pages (ASP), and the open source PHP platform.

The rest of this section describes some of the technologies in common use for developing Web software. This is a rapidly evolving field, thus any such description is automatically out of date. The first discussion is an overview of some of the client-side software technologies, then a variety of server-side technologies. The original Web server-side technology, CGI, is discussed first, followed by the relatively established J2EE platform, then the newer .NET platform. The section closes with some discussions of data handling, including XML and access to databases.

This chapter does not address Web services, which are built on top of the technologies described here. Web services (sometimes called application services) are services (usually including some combination of programming and data but possibly including human resources as well) that are made available from a business's Web server for Web users or other Web-connected programs (TechTarget, 2003). Providers of Web services are generally known as application service providers. Web services range from such major services as storage management and customer relationship management (CRM) down to much more limited services such as the furnishing of a stock quote and the checking of bids for an auction item. The accelerating creation and availability of these services is a major Web trend.

Client-Side Technologies

There are many plug-ins that Web browsers can contain to support dynamic execution. The browser is the host that supports the technology, and the plug-ins have the ability to execute certain languages and support non-textual media applications such as images, Flash, video, and sound. This is generally associated with dynamic HTML. Dynamic HTML allows client-side processing to be done by using scripting languages. Scripting languages include JavaScript, VBScript, and Jscript, all of which have similar functionalities. When used on the client-side, they can access information about the client's browser, operating environment, and hardware configuration, and access and modify information in the current Web page, and respond to user events. They cannot access server-side data when used as a client-side plug-in (although some of these scripts are also used on the server-side).

Common Gateway Interface (CGI)

One of the first technologies to perform processing on the Web server was the common gateway interface (CGI) protocol. CGI allows data to be sent from HTML form fields on the client to the server and provides a mechanism for

processing that data (server-side processing) and then returning information, usually in the form of a Web page, to the client. CGI programming allows the server to access files and other resources on the server. Although CGI is general enough to allow any programming language to be used, the most common language has been the interpretive language Perl, a very flexible scripting language that is strong in text-handling and accessing system functions.

Developers quickly found a number of limitations of CGI. Each execution of traditional CGI modules requires a new process to be created on the Web server, which severely affects performance. It has no built-in session management services, which makes it difficult to develop e-business applications. Most CGI applications have traditionally used interpretive languages such as Perl, which suffer from a number of software engineering disadvantages; in particular, most do not have capabilities such as type checking and exception handling and offer limited or no support for information hiding and inheritance. Although not a serious limitation for small applications, this makes it hard to write large Web applications that satisfy quality requirements such as reliability, usability, security, and scalability. The Apache server now includes "mod-perl" and "mod.php," which use threads to ameliorate the performance problem, but the other issues remain. One common strategy is to build an initial version of the application in CGI, either a prototype or Version 1 application, then to rewrite the application in compiled modules and scripted pages.

The J2EE Platform

Although many applications are built using CGI, the current trend is toward integrated technologies that avoid some of the disadvantages of CGI. Many of the heavy transactional-based Web sites, particularly those supporting e-business, are building new Web sites with the J2EE platform. The J2EE platform is not a product but a standard that defines the behavior of various pieces of technologies, and there are several implementations of the standard (Patzer, 2000). The standard is defined by one company (Sun), but products are available from dozens of companies, including open-source solutions. The J2EE platform, often in conjunction with Web service platforms, is currently used by many major Web-based companies and services, including well-known sites such as Netflix, eBay, Siemens, Amazon, the National Science Foundation, Major League Baseball (mlb.com), and MovieFone. This chapter discusses the individual technologies.

The J2EE platform is centered around one language, Java. Java program components are compiled to an intermediate form called "bytecode," which is executed by a Java Virtual Machine (JVM). Java bytecode is intended to be independent of hardware, operating system, and browser, and thus can be moved between computers. Java has simple built-in support for interfacing with other languages, thus providing support for connecting with legacy systems.

The primary mechanism for server-side processing in the J2EE standard is the Java servlet. Java servlets are compiled-modules that collect data from the client's Web browser into an object (the request object) with a simple API that can be accessed by servlets, and output from servlets can be returned to the client (through the response object). Servlets are Java classes that inherit from the servlet base class, and execute as lightweight threads within a plug-in called a servlet container. The container cooperates with the Web server and takes care of issues such as instantiating and destroying servlet objects, putting data from the client into the request object and returning data from the response object to the data.

The J2EE scripted page technology is Java Server Pages. A simplistic view of JSPs is as an "inside-out" version of servlets. Instead of Java classes that produce HTML, a JSP is an HTML page that includes Java statements. JSPs are first translated into Java servlet classes then compiled and run as servlets. This makes JSP execution clean and efficient; the Web server does not need a completely new plug-in module to support JSPs. In addition to the HTML, JSPs contain declarations, which are translated to Java class level variables and methods, Java scriptlets, which are translated to blocks of Java statements and that can make external method calls, and expressions, which are values printed inside the HTML.

Integral parts of the J2EE environment are Java Beans and Enterprise Java Beans. A Java bean is a design convention rather than a language feature or plug-in technology and is intended to be used to produce reusable software components. A bean is a Java class that has three characteristics: (a) it is a public class, (b) it has a public constructor that has no arguments, and (c) it has public get() and set() methods. Beans are based on the concept of a property, which is a simple data object (such as a variable) that defines some attribute of the software application. Properties should be associated with only two types of methods, getters, which return the property's value, and setters, which changes the property's value. The usual convention is that a property with name propName is accessed through the methods getPropName() and setPropName().

Despite the name, Enterprise Java Beans (EJBs) differ significantly from Java Beans. EJBs are intended to implement all of the required business logic for Web applications. They are Java classes that follow a well-defined set of rules and conventions that allow them to be installed into and executed within the confines of an EJB container. EJB containers are plug-ins that provide critical services to their EJBs. Specifically, they handle life-cycle and resource management, transaction management, data persistence, and security.

The final crucial element of the J2EE platform is the ability to conveniently interact with databases. The Java Database Connectivity (JDBC) API allows Java programs to store data into sequential databases using commands that are independent of database vendor and hardware–software platform. This allows a program to be moved from, for example, a Unix platform using Oracle's database to a Windows computer using MS Access with only a minimal number of changes. The runtime execution environment (JVM) translates the generic database statements in the program to the vendor-specific database access calls.

The .NET Platform

The Windows .NET platform collection of technologies was introduced by Microsoft as an alternative platform for building Web software applications. Its goals and much of the details of the technologies are similar to the J2EE platform, and comparatively speaking, it is conceptually easy for developers to move between the two platforms (although there are large differences in the syntax and terminology that describes the concepts) (Vawter, 2001). Microsoft .NET is partially based on the older Windows DNA, a previous Microsoft platform for developing Web applications. Windows .NET includes many technologies that were already being used, including Microsoft Transaction Server (MTS) and COM+, Microsoft Message Queue (MSMQ), and the Microsoft SQL Server database. The .NET platform includes these technologies as is or in modified forms and adds a Web services layer on top.

Whereas the J2EE platform is based on the Java programming language, .NET is intended to be language-independent and is designed to allow components in multiple languages to interoperate. Software components within .NET can be written in languages such as VB .NET (Visual Basic for .NET) or C#. C# is Microsoft's new object-oriented programming language and is very similar to Java. C# programs are first translated into Microsoft Intermediate Language (MSIL or IL). The IL code independent of platform and is analogous to Java bytecode. One key (only partially realized as of this writing) is for the IL to be independent of language, thus multiple languages can be translated into IL. If a translator is available to translate a specific language to IL, the language is called .NET enabled. The IL code is how .NET allows integration with legacy software.

The .NET platform handles server-side processing in a variety of languages, although the dominant language is currently ASP .NET. Compiled modules are translated into the IL and processed efficiently with a .NET server. Traditional ASP was a scripted page technology and is still the technology used within the .NET platform. The .NET platform includes specialized components written at the middle-tier layer, called managed components. The managed components are supported by COM+, C#, or another .NET enabled language and are used to implement business logic. Database interaction is through the ADO .NET interface.

A number of articles have compared J2EE with .NET (Farley, 2000; Middleware, 2002; Sessions, 2001; Vawter, 2001). They are all informative but the perceptive reader must take care to check the publisher, underwriter, or author for bias. Although the referenced articles should help the interested reader see more details, the differences can be summed up succinctly. As reported by Farley (2000), the cliché is that "J2EE is language-specific and platform-independent, and .NET is language-independent and platform-specific." This cliché is only half true because J2EE applications can and do include multiple languages (although most J2EE developers try to avoid multiple languages for sound engineering reasons) and many J2EE applications are restricted to single platforms. Additionally, most .NET applications use C# and the other built-in technologies so the language-independence has not, as yet, been widely taken advantage of. Being newer, .NET has also improved on some of the technical weaknesses of J2EE, including better XML support and simpler deployment.

XML as the Glue

A problem that software engineers have faced for many years is that of passing data among software components. The two components must agree on format, types, and organization. Web software applications have two unique requirements for data passing, loose coupling and dynamic integration. The fact that the components are very loosely coupled makes it more difficult for developers to establish a priori standards. The developers may be separated by time and geography and be in separate, even competing, companies. Web software applications also use dynamic integration, which means that the software engineers may not know which components will interact when the software is written.

In the 1970s, data were usually stored as records in files and the file formats were often not documented. If a new program needed to read a file, the software engineer had to deduce the file format by reading the source of the original program if it was available. If not, the engineer would usually induce the file format by trial and error—writing programs to read and print strings from the file. In the 1980s, data were usually stored in memory as abstract data types. They were saved in long-term storage in files, and both the file input–output and access to the abstract data type was managed by wrapper modules. Although much improved over previous methods, this method was usually slow, the developers of the programs had to agree on the data format, types and organization, and maintenance was often challenging because it was not clear who owned the wrapper module. These problems are exacerbated with Web software because of the extremely loose coupling, dynamic integration, and heavy reuse and use of third-party software components.

A solution from the World Wide Web Consortium (2000) is XML, or extensible markup language. XML allows data to be transferred between software components in a way that is independent of type, self-documenting, has an easy-to-understand format, and that can be parsed in simple ways. XML stores data as plain text (UNICODE) strings. Each string value is stored in between tags that are meant to imply some semantics for the contents. For example, the title for an encyclopedia article might be encoded as `<Title>Software Design and Implementation in the Web Environment</Title>`. This allows XML to be used as the primary way to pass data back and forth between Web-based software components. The principal syntax rules (Sall, 2002) are as follows:

- The document must have a consistent, well-defined structure,
- All attribute values must be quoted (single or double quotes `<Title Type = "article">`, not `<Title Type = article>`),
- White space in content, including line breaks, is significant,

- All start tags must have corresponding end tags (`"</Title>"`),
- There must be a single root element, which must contain all other elements,
- Elements may be nested, but they must not cross (`<Title>...<Name>...</Title>...</Name>` is not allowed),
- Each element, except the root element, must be contained by exactly one parent element,
- Element and attribute names are case-sensitive (`<TITLE> is different from <Title>`),
- Keywords and document type definition (DTD) elements must be all uppercase (`DOCTYPE, ENTITY, ELEMENT, and ATTLIST`), and
- Empty element tags must end in `"/>" (<editor/>)`.

Database Connectivity

A natural desire for Web software designers is to store data in general database engines. This offers a general solution to the problem of storing data and allows developers to rely on the many advantages offered by a database, including general access, efficiency of storage and retrieval, and security. Both the J2EE and .NET platforms provide general cross-vendor connectivity and data access across relational databases from different vendors. The platforms' API support mechanisms provide a convenient way to make generalized database calls from within software. The calls are made by embedding structured query language (SQL) statements into programming statements. Within J2EE, the Java Virtual Machine (JVM) uses a special JDBC driver to translate generalized JDBC calls into vendor-specific database calls. With .NET, programs connect to databases using services that Microsoft Host Integration Server 2000 provides, such as the Component Object Model (COM) Transaction Integrator (COM TI). In both platforms, programs can connect to external databases using Web services technologies such as Component Object Model (SOAP); Universal Description, Discovery, and Integration (UDDI); and Web services description language (WSDL).

The typical procedure for Web application programs to connect to databases is to start by loading the database driver. This generally has some database vendor-specific aspects and includes information about where the database is located. The second step is usually to obtain a connection to the database. Again, this requires some vendor-specific information, including security protocols (user IDs and passwords). Subsequent steps are generally completely independent of the database vendor or platform. The program should be able to create and execute database statements and then use the results (called "result sets") from statements to access data returned from the database. An obvious advantage of this approach is that the programmers do not need to know much about databases. Another advantage is that the database does not have to be local but can be anywhere on the Web (although in practice, the database is usually connected to the Web or applications servers through a secure intranet).

DESIGNING WEB SITE SOFTWARE

As a field, we are still learning how to design Web software applications. Some companies rely on prebuilt Web application servers (sometimes called "Web service platforms") such as IBM's WebSphere, BEA's WebLogic, and the open source Java Struts. Other companies buy general-purpose Web sites from vendors, which are then customized to their needs. Still others build their own Web sites completely, because they cannot afford the expensive Web service packages, because their Web sites are small enough not to need that much support, or because their needs are specialized enough so that the Web services do not support them. A complete description of how to define Web site software is certainly beyond the scope of this chapter, and, at present time, probably impossible. Nevertheless, a few hints and design strategies have emerged as being useful.

As with any software product, a crucial step is to establish a strong software requirements baseline, which should be followed by a carefully considered information architecture specification. This should include a site map, navigation among Web pages, compositions, labeling, and data element mappings. The navigation is one of the key components of usability and the literature does not contain much help for how to do this part of the design. A careful Web application design will include a high-level software design, software architecture and system architecture diagrams, class diagrams, sequence diagrams, and class specifications.

One of the most commonly used design structures is the model-view controller (MVC) architecture (Kassem et al., 2000). It provides a way to divide the responsibilities of objects. The intent is to decrease coupling between objects and layers, which supports maintenance. An MVC Web application contains three components, the model, the view, and the controller. The model encapsulates the application state, responds to state queries, presents application functionality to the user, and notifies the view of changes. The view renders the models on screen, requests updates from models, sends user inputs to the controller, and allows the controller to select a view. The controller defines application behavior, maps user actions to model updates, and selects a view to show to the user. Many other architecture styles are currently being developed.

We are also beginning to see techniques for formal modeling of Web site software applications. Sun, Song, Liu, and Wang (2001) presented an XML/XSL approach for developing Web software applications using the formal specification language Object-Z. XSL Transformations (XSLTs) are used to develop projection techniques and tools from Object-Z (in XML) and UML (in XMI). This provides a formal approach to modeling Web applications, which is not only helpful for standard e-business Web applications, but may be necessary for the Semantic Web (Berners-Lee, 1999).

Sound Software Design and Implementation Practice for Web Software

Most software in use today only has had to satisfy modest reliability requirements. The user base for Web software

is large, and users expect the Web applications to work as reliably as purchases at a grocery store or phone orders from a catalog. Moreover, if a Web application does not work well, the users do not have to drive farther to reach another store; they simply have to point their browser to a different Web site. Thus, if Web software is unreliable, Web sites that depend on the software will lose customers and the businesses may lose money. Careful use of sound development processes, full-strength languages such as Java and C#, debugging and testing tools, and well-validated third-party software components have dramatically improved reliability for some Web applications. At this point in time, however, there are many unanswered questions, including what processes succeed, how best to design the software, and how best to test the software.

Several general principles must be followed to ensure quality design of Web software. The software has to work well every time, and it must be easy to maintain, thus the design specifications must be well documented and the program must be well commented. Because the work environments tend to be dynamic and diverse, software components must be integrated, the development team must collaborate heavily, and everybody on the team should have a clear understanding of the design. Web applications need to be scalable and will change often and frequently, so many engineers believe the software must also be written to allow for future requirements. Other engineers have the opposite idea, believing that when requirements change the systems should be rebuilt from the beginning. Although this view has attracted some attention, it runs directly counter to more than 30 years of software engineering wisdom.

Web sites must be usable, so a successful development team must include one or more usability experts (Nielsen, 2000). In addition, actual users of the application must be involved with the user interface portion of the design from the beginning of the project.

CURRENT ISSUES WITH DESIGNING WEB SOFTWARE

The high-quality requirements that Web software must exhibit bring new and interesting challenges to Web software developers. This section identifies a few of these challenges; as of this writing, research is underway to develop ways to ensure the quality of software that is used for Web applications.

Design Challenges

Tremendous effort has been expended to ensure the quality of traditional programs, resulting in testing techniques for both stand-alone and distributed systems. Although some of these techniques can be used to help ensure the quality of Web applications, some of the special features and requirements of Web applications prevent them from being directly adopted. These challenges are summarized in the following paragraphs.

The overall architecture of Web applications is similar to client–server systems in many aspects, but there is a key difference. In traditional client–server systems, the respective roles of the clients and servers and their interactions are predefined and static. In Web applications, however, client-side programs and contents may be generated dynamically. For example, a server may return a dynamically generated HTML file that contains dynamically generated JavaScripts, links and contents. This means that which subsequent interactions between the client and server are available depend on the previous inputs.

For traditional programs, correctness and efficiency are usually the most important quality factors. For Web applications, other quality features can often be more important, and yet we have few techniques for supporting them. For example, compatibility and interoperability are urgent and cause problems that are more serious than with traditional programs. Traditional programs are usually developed for a certain predefined, well-understood environment, with few conflicts and changes. Web applications are often affected by factors that may cause incompatibility and interoperability issues. For example, server components can be distributed to different operating systems, such as UNIX, Linux, Windows, MacOS, and AIX, each of which has multiple versions, and run with different Web server packages, including IIS from Microsoft, Apache from the Apache software foundation, WebLogic from BEA, WebSphere from IBM, and others. The situation is even more complex on the client side, with different versions of Web browsers running under a variety of operating systems. Clients may also use different connection approaches, such as dial-up modems, direct Internet access, or wireless, and they may also use different Internet service providers. All of this heterogeneity makes it more difficult to produce Web application components that are compatible with one another and that interoperate easily and correctly.

Another difference between Web applications and other types of programs is the variance in the control of execution of the application. For traditional programs, the control flow is fully managed by the program, so the user cannot affect it. When executing Web applications, users can break the normal control flow without alerting the program controller. For example, users can press the back or refresh button in the Web browser, which changes the execution context, causing unexpected results. Furthermore, changes in the client-side configuration may affect the behavior of Web applications in ways that are difficult for Web software designers to anticipate. For example, users can turn off cookies, causing subsequent operations to malfunction.

Web applications also have much faster maintenance requirements than most traditional software. Web technologies evolve more rapidly than traditional software technologies, and the changes in Web application requirements can be more dramatic—maintenance not only needs to be done more frequently, but more efficiently.

Web applications also have features that are not present in client–server and distributed systems. These include session control, cookies, the stateless aspect of HTTP, and new security issues (related to the use of public networks). Therefore, new solutions are necessary to implement these features correctly.

GLOSSARY

Many of the definitions in this glossary are derived in whole or part from TechTarget's definitions, including "whatis.com" (TechTarget, 2003). More details can be found on their Web site for some of these terms.

Active Server Pages (ASPs) A scripted page technology that uses HTML templates that can include programming statements. ASPs predated Microsoft's .NET but have been folded into the platform.

Application program interface (API) The specific method prescribed by a computer operating system or by an application program by which a programmer writing an application program can make requests of the operating system or another application.

Application server A server program in a computer in a distributed network that provides the business logic for an application program. The term is sometimes used to refer to the software, sometimes the hardware, and sometimes both.

Browser extensions A compiled program that is written to a browser API, usually for extending the capability of a client browser to play new media forms such as audio or video. For Netscape browsers, such programs are dubbed plug-ins. Internet Explorer browsers use ActiveX controls and other kinds of plug-ins. For the J2EE platform, the equivalent is a Java applet.

Bytecode An intermediate language, similar to computer object code but usually at a higher level of abstraction. It is interpreted by a program, usually referred to as a virtual machine, rather than by the actual hardware. Java is translated to a bytecode that is optimized for fast interpretation that can be executed on a number of platforms by the Java virtual machine.

C# (pronounced "C-sharp") An object-oriented programming language from Microsoft that combines elements of C++ with Visual Basic. C# is has many features in common with Java.

Common gateway interface (CGI) A protocol that defines how data is sent back and forth between Web clients and external server-side programs. Input in CGI comes from HTML form data and HTTP headers (termed environment variables) and output set by HTTP headers indicating multi-purpose Internet mail extensions (MIME) type and common Web formats such as HTML.

Client–server computing A model of computing in which one computer or software component (the server) manages and provides access to resources to another (the client) by responding to requests.

COM+ The .NET middle-tier infrastructure designed to support business components.

Compiled (Web-server) modules A compiled program that that is built into a Web server API such as Apache Modules or Microsoft IIS Internet server application program interface filters or modules. Input and output with server modules is similar to CGI programs but generally is much faster and happens at a much lower level. For the J2EE platform, such modules are dubbed servlets.

Cookies A text string that a Web application stores on a client through the client's Web browser. The intent is to use the string as an index to retrieve information about the user who is associated with the cookie, thereby keeping track of state information that is passed between a server and the user.

Dynamic HTML A collective term for HTML tags and options that support animation user interaction. The tags and options include the ability to respond to user events client-side using BOM–DOM (browser object model–document object model) and scripting languages such as JavaScript.

E-business A company that does all or an important part of its business over the Internet.

Enterprise Java Beans (EJB) Java classes that follow a well-defined set of rules and conventions that allow them to be installed into and executed within an EJB container, which provides services such as life-cycle and resource management, transaction management, data persistence, and security.

Hypertext markup language (HTML) The most common language used to create Web pages.

Hypertext transfer protocol The fundamental network protocol that Web browsers and servers use to communicate. It is a lightweight, connectionless protocol.

Intermediary language (IL) The intermediate language used by the.NET platform.

The Java 2 Enterprise Edition (J2EEE) platform A collection of conventions, plug-ins, and library packages that support Web software. It includes Java servlets, JSPs, Java beans, and EJBs.

Java A general purpose object-oriented programming language. Java is extended by libraries that contain packages and code that support Web software.

Java Server Pages A scripted page technology that uses HTML templates that can include Java statements. Java Server Pages are first translated into Java servlet classes and then compiled and run as servlets.

Java Applets A Java class that can be included in an HTML page. The Java bytecode is transferred to the client's computer and then executed by the browser's Java Virtual Machine (JVM). One common use of applets is to produce high functionality GUIs.

Java Beans A Java class that is used to create reusable software components. A Java Bean is expected to have three characteristics: (1) it is a public class, (b) it has a public constructor that has no arguments, and (c) it has public methods to assign and retrieve values of objects called properties. By convention, the methods are called get() and set().

Java Data Base Connectivity (JDBC) An application program interface (API) specification for connecting Java programs to common databases. Database commands in SQL are embedded in Java programming statements and the API handles most of the interaction invisibly.

Java Servlets A compiled module technology; a Java class that inherits from the servlet base class and executes as lightweight threads within a plug-in called a servlet container. Servlets run on the server, accept requests from the Web server, and generate responses for the client, usually in the form of HTML pages.

JavaScript The common name of a Web scripting language based on ECMAScript, which is unrelated to Java in all but name. JavaScript is traditionally used within Web browsers for validation of form data and other basic tasks but with the rise of complex document object modules, JavaScript is increasingly being used to perform complex client-side manipulations. The use of JavaScript in such a fashion is a major part of dynamic HTML or DHTML.

Middleware Layers of software between client, server, and other N-tier levels that provide services such as communication. Often bought from a third party vendor.

N-tier architecture A software architectural design with components that are broken up into two or more (N) layers, where each layer only communicates with its two adjacent layers.

.NET A collection of conventions, plug-ins, and library packages that support Web software on Microsoft platforms.

Plug-ins Programs that are installed and used in the context of a Web browser; used to process particular types of files from the Web server such as PDF, Flash, and Java.

Scripted pages HTML templates that process business logic by executing on the server side, not client side, and that can access software on the server to get and process data. Common server-side scripting environments include Active Server Pages (ASP), ASP.NET, ColdFusion, Java Server Pages (JSP), and the PHP platform.

World Wide Web Consortium An organization with the responsibility of leading the development of the Web, its technologies, and its standards.

Web site engineering The application of well-documented principles, techniques, and technologies to develop software for the Web that is of high quality, where the quality must satisfy goals in terms of measurable criteria such as reliability, usability, security, availability, scalability, maintainability, and performance.

Web server A program that supplies Web pages and other Web services to clients, or a computer that makes software available through Web protocols.

XML (extensible markup language) A flexible way to create common information formats and share both the format and the data among programs.

CROSS REFERENCES

See *Client/Server Computing; Common Gateway Interface (CGI) Scripts; DHTML (Dynamic HyperText Markup Language); Extensible Markup Language (XML); HTML / XHTML (HyperText Markup Language/Extensible HyperText Markup Language); Java; Java Server Pages (JSP); JavaBeans and Software Architecture; JavaScript; Middleware; Web Site Design.*

REFERENCES

Berners-Lee, T. (1999). *Weaving the Web*. San Francisco: Harper.

Constantine, L. L., & Lockwood, L. A. D. (2000). Software for use: A practical guide to the models and methods of usage centered design. ACM Press.

Carnegie Mellon Software Engineering Institute (2002). Capability maturity model for software (SWE-CMM). Retrieved April 24, 2002, from www.sei.cmu.edu/cmm/

Dustin, E., & Rashka, J., & McDarmid, D. (2001). Quality Web systems: Performance, security, and usability. Addison-Wesley.

Farley, J. (2000). Microsoft .NET vs. J2EE: How do they stack up? Sebastopol, CA: O'Reilly & Associates. Retrieved August 1, 2000, from http://java.oreilly.com/news/farley_0800.html

Kassem, N., & the Enterprise Team. (2000). *Designing Enterprise applications with the Java 2 platform (Enterprise Edition)*. Boston, MA: Addison-Wesley.

Menascé, D. A. (2000). *Scaling for e-business: Technologies, models, performance, and capacity planning*. Upper Saddle River, NJ: Prentice Hall.

Middleware Company. (2002). The Petstore revisited: J2EE vs .NET application server performance benchmark. Retrieved October 2002 from http://www.middleware-company.com/j2eedotnetbench/

Murugesan, S., & Deshpande, Y. (2001). Web engineering: A new discipline for development of Web-based systems. In S. Murugesan & Y. Deshpande (Eds.), *WebEngineering 2001* (pp. 3–13). Berlin, Germany: Springer-Verlag Lecture Notes in Computer Science 2016.

Nielsen, J. (2000). *Designing Web usability*. Indianapolis, IN: New Riders.

Offutt, J. (2002). Quality attributes of web software applications. *IEEE Software* [Special issue on software engineering of Internet software], *19*(2), 25–32.

Patzer, A. (2000). *Professional Java Server programming (J2EE edition)*. Chicago, IL: Wrox Press.

Powell, T. A. (1998). *Web site engineering: Beyond Web page design*. Upper Saddle River, NJ: Prentice Hall.

President's Information Technology Advisory Committee (1999). Information technology research: Investing in our future (Technical Report). Washington, DC: National Coordination Office for Computing, Information, and Communications. Retrieved February 7, 2003, from www.ccic.gov/ac/report

Sall, K. (2002). XML syntax rules, Web developers virtual library. Retrieved April 5, 2002, from http://www.wdvl.com/Authoring/Languages/XML/XMLFamily/XMLSyntax/sall2_1.html

Scharl, A. (2000). *Evolutionary Web development*. Berlin, Germany: Springer.

Schneider, F. B. (1999). Trust in cyberspace. Washington, DC: National Academy Press. Retrieved February 7, 2003, from http://www.nap.edu/html/trust/

Sessions, R. (2001). Java 2 Enterprise edition (J2EE) versus The .NET platform: Two visions for eBusiness. ObjectWatch. Retrieved March 28, 2001, from www.objectwatch.com/FinalJ2EEandDotNet.doc

Sun, J., Song, J. D., Liu, J., & Wang, H. (2001, May). *Object-Z Web environment and projections to UML*. Paper presented at the Tenth International Conference on the World Wide Web, Hong Kong, China.

Tassey, G. (2002, May). The economic impacts of inadequate infrastructure for software testing (Research Triangle Institute, NIST Technical Report 7007.011). Retrieved February 27, 2003, from http://www.nist.gov/director/prog-ofc/report02—3.pdf

TechTarget (2003). Whatis.com, part of the TechTarget family of Enterprise IT Web sites. Retrieved February 27, 2003, from http://whatis.techtarget.com/

World Wide Web Consortium (2000, October). Extensible (XML) 1.0 (2nd ed.). W3C recommendation (W3C #28). Retrieved February 7, 2003, http://www.w3.org/XML/#9802xml10

Vawter, C., & Roman, E. (2001). J2EE vs. Microsoft .NET: A comparison of building XML-based Web services. Retrieved June 2001 from http://www.theserverside.com//resources/article.jsp?l = J2EE-vs-DOTNET

Software Piracy

Robert K. Moniot, *Fordham University*

INTRODUCTION

"Don't copy that floppy!" is the rallying cry of the software publishers' organizations. Perhaps nowadays the slogan needs to be updated to "Don't copy those warez!" as the focus of activity shifts to the Internet. But whatever form it takes, there is no question that piracy is a major problem facing the software industry. In the year 2001, an estimated 40% of all copies of business software applications installed worldwide were pirated, having a retail value of some $11 billion. This lost revenue deprives software companies of the remuneration to which they are entitled for their efforts in developing and distributing software. It potentially may increase prices for software and inhibit innovation of new products and may also cause some companies to go out of business.

MODES OF SOFTWARE PIRACY

Software piracy is any copying of software in contravention of its license. One of the biggest obstacles to reducing piracy is the widespread ignorance of what actions constitute piracy. Here are some ways that piracy can occur:

Downloading proprietary software from an unauthorized Internet bulletin board or Web site, or directly from another user via a peer-to-peer file sharing program.

Purchasing counterfeit software in a store or at an Internet Web site or auction.

Borrowing the medium containing an application purchased by an employer for use at one's place of work and installing it on a personal computer at home.

Borrowing a program from a friend, a co-worker, or a library, and installing it on one's own computer.

Selling or giving away an old version of a program after receiving an upgrade.

Leaving an installed program on an old computer after installing it on a new computer without purchasing a new copy of the program.

Installing more copies of a program on the computers in an enterprise than the license allows, or installing it on a server for use over a local area network if this is not permitted by the license.

Note that it is always permissible to make a copy of software for backup or archival purposes, but any such copy must be destroyed if the user no longer can legitimately use the program. Also, users may sell or give away programs they legitimately own to someone else, provided they do not retain their copies. For instance, users can leave installed software on old machines that they sell or give away if they purchase new computers with new software preinstalled.

The term "piracy" has long been used to mean acts of infringement of copyright. Thus in recent times it was natural to adopt the term to include the illicit copying of software, even before the application of copyright law to software was fully clarified. However, piracy is a broad term encompassing many diverse forms of infringement, only some of which are listed above. Each of these forms has its own legal and ethical ramifications, as well as distinct perceptions by its practitioners. One important distinction is between copying for private use only, or end-user piracy, and copying for sale. Many people consider copying for personal use as either acceptable or having only minor ethical significance, whereas most recognize copying for sale as both unethical and illegal. Another distinction is between small-scale and large-scale piracy. Although each act of small-scale piracy is relatively minor, the aggregate effect is quite large. In fact, small-scale copying for personal or corporate use is said to be the most widespread form in practice and to account for over half the total value of pirated software (Software and Information Industry Association, 2000). The growth of the Internet as a medium for exchange of software has greatly facilitated this form of piracy.

End-User Piracy

Small-scale piracy mainly takes the form of "softlifting," which means copying by individuals for their own personal use. Softlifting can be done in a wide variety of ways. Probably the most common method is to borrow the installation media from a friend or co-worker. Or instead of

borrowing the original media, one might obtain an unauthorized, or "bootleg," copy. Bootlegging by sharing of software over the Internet is also frequent. Before the advent of the World Wide Web, individuals often posted software on Usenet newsgroups or on bulletin boards. Nowadays there are thousands of Web sites that post "warez," or contraband software, for download. More recently, peer-to-peer systems have been developed that allow individuals to share software with each other directly.

Renting software and not uninstalling it after use was once a fairly common mode of softlifting. For this reason, the unauthorized renting of software was made illegal in the United States in 1990. Web sites offering software rental can be found on the Internet, but it does not seem that this is a prevalent mode of softlifting nowadays. The law permits libraries to lend software, provided that the package contains a clear copyright notice. Quite likely these loans are often used for softlifting.

Closely related to softlifting is "softloading," or the installation of a legitimately purchased program onto more machines than the software is licensed for. It can also involve the installation of the software onto a server for use by multiple client machines in a local area network. Softloading usually occurs in a corporate setting, which can be a business, a nonprofit institution such as a university or hospital, or a government agency. It can occur inadvertently, if the information technology staff does not keep proper records of licenses and the number of installed copies of each software application.

Commercial Piracy

Industrial piracy can take two very different forms: counterfeiting and cloning. Counterfeiting is the reproduction of packaged software for sale. Sometimes the counterfeiting is done in such a way as to make it appear to be authentic, so that it can be sold for a price that is comparable to the normal retail price. These counterfeiters take care to duplicate the appearance of the media, the packaging, and even the documentation as closely as possible. The purchaser may be unaware that the item is not genuine and will be unpleasantly surprised to find it is not entitled to support such as upgrades from the manufacturer. There may be telltale indications of piracy, such as poorly reproduced artwork, misplaced logos, misspellings, or a missing authenticity hologram. In other cases, the counterfeiters make no attempt to conceal the pirated status of the product, and it is sold for an extremely low price. This practice is also called bootlegging. Often a number of bootleg applications with a market value of hundreds of dollars are bundled together on a single CD that may sell for $20 or less.

Cloning is the independent creation of a functional duplicate of an existing program, which is typically marketed as an independent product. An example was the case of Paperback Software's VP-Planner, which closely imitated the functionality and user interface of Lotus Development's popular spreadsheet program 1-2-3. Cloning takes considerable programming effort, but avoids the laborious prototyping and design effort involved in the creation of a totally new program.

Counterfeiting and cloning are the easiest forms of piracy for software producers to combat, provided there is support from the authorities in the host country. This is because they most closely resemble traditional forms of copyright or patent infringement, for which legal remedies are well established. Furthermore, the offender is often readily identified, and a lawsuit is likely to yield a substantial return in the form of damages and penalties.

Original equipment manufacturers (OEMs) produce personal computers that are typically sold fully loaded with an operating system and a suite of applications. The OEMs typically enter into licensing agreements with the software producers to authorize the installation of this software. OEMs or hardware dealers sometimes illegally load software onto more machines than authorized, or they may load software that was not included in the license agreement, as a way of making the computers more attractive for sale. This practice is called "hard-disk loading." "Unbundling" is the sale of OEM-version software items separately from the computer system for which they are authorized. "Mischanneling" is the diversion of specially discounted software, intended for academic institutions, government agencies, and other high-volume customers, for sale to others who do not qualify for these discounts.

MOTIVATIONS FOR SOFTWARE THEFT

Why does an individual choose to steal software? On the other hand, if obtaining an illicit copy of a software application is so easy and cheap, why does anyone purchase the legitimate article? Probably the reader can think of several likely motivations on either side, but a number of studies have been done in an effort to provide well-founded answers to these questions. (See, for instance, Cheng, Sims, & Teegen, 1997; Simpson, Banerjee, & Simpson, 1994; Taylor & Shim, 1993.) Most of these studies have been based on surveys of students and business executives. These studies are not always directly comparable, because they take different approaches and use different models of softlifting attitudes and intentions. They also vary in the way they validate the measures used and control for various biases. Furthermore, it is possible that some of the reasons given may be rationalizations rather than true motives. Despite these limitations, some consistent patterns emerge from these studies.

Probably the most important conclusion is that the primary reasons for softlifting are economic: the software is seen as overpriced, or the individuals cannot afford it. Another common reason is the desire to try out the software before buying it, or to use it for only a short time. On the other hand, individuals are more likely to purchase the software if they feel that it will be useful for schoolwork or on the job and if it will be frequently used. Another motive for purchasing is the availability of user manuals and technical support. A significant finding of the studies is that the perception of softlifting as unethical, illegal, or against school or company policy has little effect on the decision to softlift. However, a perception that softlifting is acceptable and prevalent among one's peers increases the likelihood of softlifting.

Other studies have tried to identify cultural and socioeconomic indicators that are predictors of software piracy rates. These studies have the advantage of using software industry estimates of piracy rates rather than relying on self-reporting in surveys, which is an unreliable indicator of actual behavior. On the other hand, these studies perforce use data at the level of whole nations and so necessarily average out the differences between individuals or between regions within a given country. It should be noted that the piracy data on which these studies are based include only business software. There is probably a strong correlation between business and personal copying of software in each country, and so the results should be applicable to rates of individual softlifting as well. Marron and Steel (2000) and Husted (2000) found that lower piracy rates are associated with higher levels of economic development (per capita GDP or income), with greater disparities in income within a country (implying a smaller middle class), and with stronger institutions to enforce contracts and protect property from expropriation. They also found that individualist cultures, i.e., those that value individual rights and ownership, have lower piracy rates than more collectivist ones that put greater value on mutual help and sharing. They did not find a significant correlation with the average level of education.

These results are reasonable. Higher levels of economic development mean that individuals and businesses are more able to pay for software. In countries with greater income inequalities, the lower classes are unable or barely able to afford computers at all, and so most technology purchasing is done by the wealthy who can easily afford to pay. It is the middle classes, often struggling to make ends meet, that are the most likely to seek to cut costs by pirating software. Individualist cultures, and those with strong institutional protection of property and contract rights, are characterized by attitudes that will be less likely to view softlifting as legitimate. Collectivist cultures, in contrast, tend to deemphasize rights of individual ownership in favor of the duties of cooperation and sharing of the fruits of one's creativity for the benefit of society. Therefore those countries (which include many in southern and eastern Asia) have been reluctant to grant Western-style copyright protection to software, and even where such protection is provided by law, it must compete in the moral sphere with strongly held traditional values of community and solidarity.

Implications of the Studies

The findings of all the studies cited above carry some implications for software publishers' efforts to reduce the rates of software piracy. First, it appears that educational programs aimed at increasing individuals' awareness of the illegal and unethical nature of softlifting will be of limited effectiveness. The studies show that simple awareness of the illicitness of softlifting has little effect on behavior. Technical copy protection mechanisms (discussed in a later section) are also unlikely to be effective. They are inevitably defeated and may actually encourage piracy due to the challenge they present. On the other hand, perceived consequences, in terms of benefits as well as penalties, are important factors in most individuals' decisions whether or not to softlift. The studies indicate that increasing the likelihood of being caught and punished would deter softlifters. However, it is impractical to prosecute individual softlifters, and besides, an overly aggressive enforcement program could backfire by creating an adverse public reaction.

It appears that the most practical and effective means available to the software publishers for reducing softlifting is to lower prices (perhaps charging different categories of customers different prices) while enhancing the perceived value of products by providing user manuals, technical support, and inexpensive upgrades. The studies show that if individuals value the software for its usefulness, and value the support provided by the vendor, they will be more willing to pay for it. The validity of these reasons is confirmed by the observation that the Linux operating system and its accompanying application software from the GNU organization and elsewhere are successfully sold by a number of vendors, even though the software is all legally obtainable for free over the Internet. These vendors succeed in charging money for the software because they provide valuable support services, including documentation and help lines. Firms that depend on computer systems for their daily operations willingly pay for such support because they want to have someone to turn to for help when something fails.

Organizations That Combat Software Piracy

There are two main trade organizations that represent the software industry in its efforts to counter the illicit traffic in software. The Business Software Alliance (BSA, http://www.bsa.org) is an international organization representing major software and e-commerce developers. Its membership includes such flagship companies as Microsoft, Apple, and Adobe. Founded in 1988, its mission is to educate computer users about copyrights, to lobby for intellectual property legislation, and to combat software piracy. The Software and Information Industry Association (SIIA, htp://www.siia.net) is a coalition of software and electronic content producers. It was formed in 1999 from the merger of the Software Publishers Association (SPA, founded in 1984) and the Information Industry Association (IIA). Its membership includes some members of BSA, but also includes many smaller software and information technology companies. Its mission is to promote the interests of the software and digital information industry, to provide knowledge resources to member companies, and to fight software piracy. SIIA still uses the name SPA for its antipiracy arm.

SCOPE AND IMPACT OF PIRACY

Estimated Piracy Rates

Estimating the extent of software piracy is not a simple task. Obviously, many of the transactions whereby people obtain illicit copies of software are conducted in secrecy, and Internet warez sites do not usually keep careful records of downloads. Consequently any estimates of piracy rates must be indirect. One of the most widely cited estimates of piracy rates and of the economic impact of

piracy is produced by the International Planning and Research Corporation (IPR), a specialized consulting firm. This study has been commissioned annually since 1994 by the BSA and SIIA.

The basic quantity estimated in the IPR study is the piracy rate, defined as the number of illicit copies of software applications in use, divided by the total number of copies in use. Thus a piracy rate of zero would mean that all software was acquired legitimately, whereas a piracy rate of 100% would mean that all software was pirated. The IPR study estimates the piracy rate globally, as well as on a regional and country-by-country basis (Business Software Alliance, 2002). The IPR study considers only business software applications. Not included in the estimates are operating systems, custom software, and typical home-use applications such as recreational, educational, or personal finance software. Piracy rates for these excluded categories are probably different from the rates for business software.

The piracy rates calculated by IPR for the years 1994 to 2001 are shown in Figure 1. This chart shows a steady decline in global piracy rate from 49% in 1994 to 36% in 1999, followed by an upturn to 40% in 2001. The IPR report attributes the decline in piracy rates seen in the first part of this period to a number of factors, including the efforts made by the software industry and national governments to educate the public about copyright laws and to enforce those laws. Also, during that time, U.S. software companies made efforts to increase their presence in overseas markets, including providing better user support, while at the same time software prices generally declined. These developments made the option of purchasing software legitimately more attractive in those countries. It is too soon to tell whether the recent increases in piracy rates represent a reversal of the previous declining trend or merely a temporary fluctuation. The IPR report attributes these increases to the increased competitive pressures during a period of slower economic growth, which led businesses to be more willing to pirate software in order to cut costs. If this explanation is correct, then we may expect to see piracy rates resume their decline once worldwide economic conditions improve.

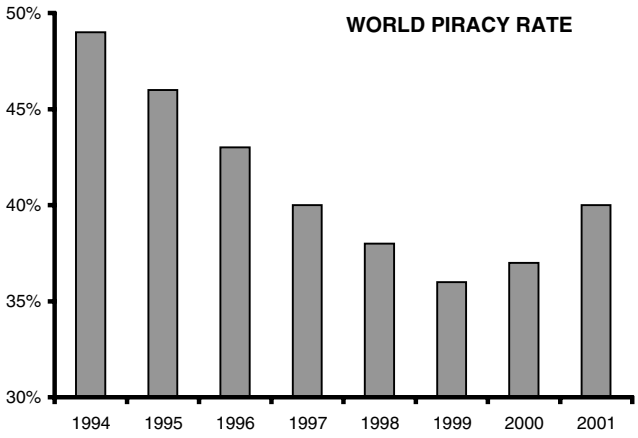


Figure 1: World average piracy rate of business software for the period covered by the IPR studies. Adapted from Business Software Alliance (2002) with permission from International Planning and Research Corp, and extended back to 1994 with data from Business Software Alliance (1999).

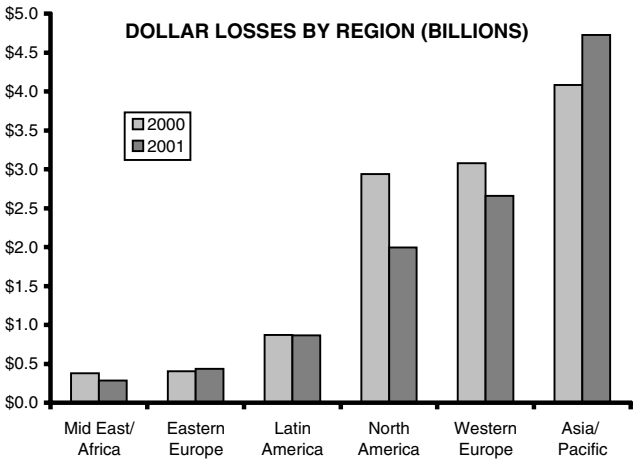


Figure 2: Market value of pirated business software by region for 2000 and 2001 (Business Software Alliance, 2002). Reproduced with permission from International Planning and Research Corp.

Figures 2 and 3 show the estimated dollar losses and piracy rates, respectively, by region for 2000 and 2001. Comparing Figure 2 with Figure 3, we see that North America and Western Europe, which have the lowest piracy rates, have some of the highest dollar losses. This seeming paradox is explained by the fact that these two regions are the largest consumers of software. A low percentage of a very large figure can be larger than a high percentage of a small figure. The data also show that the increase in the global piracy rate from 2000 to 2001 results mainly from changes in the Asia/Pacific region, where the increased piracy rate was coupled with a substantial increase in the already large amounts of software used.

Country-by-country piracy data from the IPR study are not shown here for reasons of space. Here we summarize the results in terms of a few broad patterns. First, with

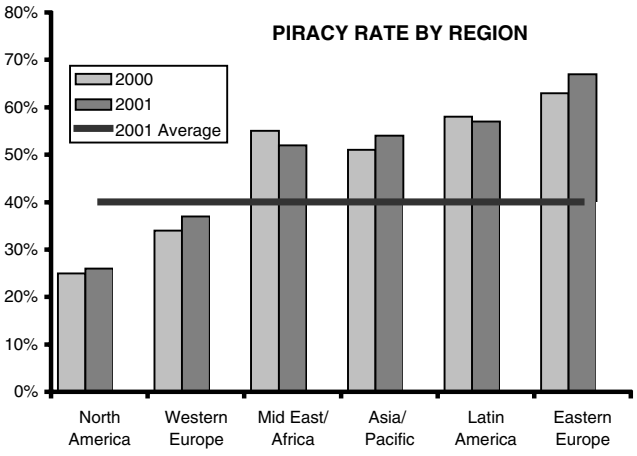


Figure 3: Piracy rate of business software by region for 2000 and 2001 (Business Software Alliance, 2002). Reproduced with permission from International Planning and Research Corp.

the exception of Eastern Europe, all of the countries with piracy rates over 70% are Third World countries. In most of these countries there is little indigenous software production, and so for them piracy has a net economic benefit without adversely impacting their own local industry. In this regard it is noteworthy that India, which has become a major player in the software industry, reduced its piracy rate significantly, to a low of 61% in 1999 from 79% in 1994 (although the rate has climbed again since then, to 70% in 2001). China's piracy rate, on the other hand, declined only slightly, to 92% in 2001 from 97% in 1994. China was seeking World Trade Organization membership during this period and was consequently under pressure to improve its protection of intellectual property rights, but its efforts in this direction evidently had little effect.

Eastern Europe has consistently had the highest piracy rate of the six regions of the IPR study for each of the years shown in Figure 1. However, if we exclude Russia and most other countries of the former Soviet Union, the rest of the Eastern European countries (including the Baltic States) have made substantial progress recently, reducing their software piracy rates by at least 15 percentage points over the period 1994–2001. Their current piracy rates still have room for improvement, ranging from a high of 75% (Bulgaria and Romania) to a low of 43% (the Czech Republic), and averaging about 54%. The large reductions in piracy rates in these countries probably reflect the emergence of free-market economies, a process that is normally accompanied by improvements in protection of intellectual property (Software and Information Industry Association, 2000). This transition has apparently not happened in the countries of the former Soviet Union, which had a combined piracy rate of 87% in 2001, only a modest decline from 95% in 1994.

Methodology of the Study

In order to calculate the piracy rate for a given country or region, the IPR needs to determine the number of pirated copies and the number of legitimately purchased copies of software applications in use in that region. The IPR has access to proprietary sales information from the BSA member companies, and so the figures for legitimate software sales are readily obtained. The number of pirated copies is inferred as the difference between this number and the "demand" for software, i.e., the number of application programs that one would expect to be loaded onto the computers in use. The IPR uses hardware sales data to determine the number of new machines sold each year. A correction is applied for turnover as older machines are taken out of service and replaced by new ones, to obtain an estimate of the number of machines currently in use. From this figure the demand is calculated based on the average number of application programs that would normally be installed on each computer. The IPR calculation of demand takes into account differences between home and business computers and different levels of technological development in different regions, as well as differences among the various categories of software.

It is important to keep in mind that much software is excluded from the IPR estimates. Only packaged (as opposed to custom or in-house developed) software is analyzed, and only packaged business software applications are included in the final piracy figures. According to PricewaterhouseCoopers (1999), packaged business software currently represents less than 20% of all software revenue. Freeware and shareware also escape the analysis because there are no sales records for them, so they will be counted among the pirated applications. Freeware is software that its producer makes available free of charge, and shareware is software that is distributed for free with a request that users who like it should send the author a contribution. At present freeware and shareware represent an insignificant fraction of the market.

Despite all of the uncertainties involved, the IPR piracy estimates are probably the best obtainable under the circumstances. Marron and Steel (2000) performed a regression of the IPR estimated piracy rates on an independently estimated measure of patent protection in various countries and found a strong correlation, as would be expected. This test gives some confidence in the basic validity of the IPR data.

Financial Impact of Piracy

The estimates of the dollar value of the revenue losses to piracy shown in Figure 2 were obtained simply by multiplying the number of pirated copies by the average market price of a copy. The software categories not included in the IPR figures probably have different piracy rates from those for packaged business software. For instance, the study by Cheng et al. (1997) suggests that piracy rates for game software are probably much higher than those for productivity applications such as word processors. In any case it is likely that piracy in the excluded categories represents a large additional amount of lost revenue. Thus, the calculated losses are probably a considerable underestimate of the total revenue lost to piracy by the software industry. At any rate, it is reasonable to assume that the ratios of the revenue losses in Figure 2 represent the relative financial impacts of piracy in different regions.

Of course, calling the revenues in Figure 2 "losses" implicitly assumes that all software now pirated would be purchased through legitimate channels at current prices if all piracy were stopped. But it is likely that some users who currently pirate software would choose not to purchase the software at all, if piracy were somehow made impossible. Also, very probably prices would change in such an altered market. Thus it would be more precise to refer to this quantity as the "market value" of the pirated software.

Economic models have been developed to consider the overall effect of illicit copying on software producers' revenues. Slive and Bernhardt (1998), among others, describe how, in some situations, piracy of a software product can actually increase the total profits of its manufacturer, through what economists call "network externalities." In essence, the value of a particular software product is increased by having a large community of users. For instance, the users enjoy the convenience of being able to interchange files in the format used by the application. Also, they may invest considerable time in learning to use the application, and can then move more easily to a new employer where the same software is in use than to one

that uses some other product. In this context, piracy can be viewed as a form of price discrimination (the practice of charging different prices to different customers) in which the software is effectively sold at zero price to some customers. The resulting increase in the size of the user community enhances the value of the software, leading to increased sales and possibly also allowing the vendor to charge higher prices for it to those customers that pay. A key element of this effect is the existence of two distinct populations of users. Home users generally place less value on software than businesses do and are less likely to be willing pay for it. Businesses value the software more, and are also likely to pay for it for the sake of reliability and support. Businesses and other organizations are also more easily targeted by antipiracy campaigns. Hence the software companies may find it in their best interest to turn a blind eye to home-use softlifting, knowing that it is helping them to build market share that pays dividends in the more lucrative business market.

For small firms attempting to enter the software market, it is unlikely that the positive network effects of piracy will be sufficiently strong to compensate for the revenue losses. This is especially the case for non-U.S. software producers. They have an inherent advantage in their home countries for producing certain types of applications such as manufacturing, banking, and financial software, areas that are the most dependent on local laws and business practices. But this advantage is defeated if they are forced to compete against extremely low-cost pirated copies of software produced in the U.S. (PricewaterhouseCoopers, 1999).

MECHANISMS FOR PROTECTION OF SOFTWARE

Introduction

Softlifting has been going on ever since there was anything to softlift. The very first consumer application produced by Microsoft was a Basic language interpreter for the Altair microcomputer that appeared in 1975. A paper tape containing a demo version of the interpreter was stolen and soon the demo was circulating widely among the Altair user community. Subsequent software products have fared no better. Industry efforts at education and persuasion have met with only limited success. Therefore the industry soon turned to legal and technical measures to protect its interests.

Legal Protection Mechanisms

Goals of Legal Protection Mechanisms

The laws protecting intellectual property rights generally have as their primary goal the fostering of a healthy creative industry for the benefit of society as a whole. For instance, Article I, Section 8 of the United States Constitution states, "The Congress shall have Power to . . . promote the Progress of Science and useful Arts, by securing for limited Times to Authors and Inventors the exclusive Right to their respective Writings and Discoveries." The notion expressed in this phrase is that individuals will be encouraged to create new inventions, writings, and other creative works if they are assured of reasonable remuneration for their efforts. Therefore the creator of a work must be granted a temporary monopoly on the production and sale of the work. Patent and copyright legislation is carefully crafted to limit the property owner's control of the creation, because a healthy creative industry also depends critically on the interchange of new ideas. The monopoly protection is made contingent on the publication of the idea or its expression, so that others can build on the idea in further creative works. Thus the goal of intellectual property protection laws is not the enrichment of the owners of the property. This is a means, not an end. The end is the provision to society of an abundant supply of high-quality inventions, writings, music, and other creative works.

Some have argued that existing legal conceptions of intellectual property are not adequate for dealing with software (Davis, Samuelson, Kapor, & Reichman, 1996). Copyright laws have traditionally been applied only to writings and other forms of expression, whereas patents were applied only to discoveries and useful inventions. But software is precisely a useful object that happens to be expressed as a writing. Notwithstanding these objections, the existing legal framework has been adapted to deal with software by a process of relatively narrowly focused changes to existing legislation accompanied by court decisions to establish precedents, just as happened with earlier technological innovations such as audio and video recorders.

Still, it should be kept in mind that this legal framework is only one of a number of possible solutions to the problem of ensuring for society an abundant supply of creative works. It is a solution that is situated within the Western tradition of protection for intellectual property. Other solutions, such as state support for software producers, or a system of shared royalty payments funded by hardware or media sales, can be imagined. The "open source" software movement promotes an alternative mode of software production in which widely dispersed programmers contribute to development of an application. However, to date none of these alternatives has found a viable mechanism for funding the work at the levels that would be required to support the current software establishment.

In what follows, the discussion will be mainly in terms of the U.S. legal situation. This is reasonable because most software is produced in the U.S. Also, various international treaties have established a legal climate that is substantially the same in most countries. These treaties include the Berne Convention for the Protection of Literary and Artistic Works, the Universal Copyright Convention, and the General Agreement on Tariffs and Trade (GATT) accord on Trade-Related Aspects of Intellectual Property Rights (TRIPs).

Software Copyright

Because a program must be copied from a distribution medium to a computer's hard drive in order to be installed, and copied again from the hard drive to the working memory (RAM) in order to run, copyright in principle gives the software maker complete control over the use of its product. On this view, the customer does not purchase the software, but only pays for a license to use it, under whatever terms the maker chooses to dictate. In practice,

the strictness of this view is tempered by market pressures and the recognition by the courts that the customer is entitled to certain basic rights. As one example, most packaged software carries a notice on the outside of the package saying that the purchaser indicates acceptance of the license by the act of opening the package. However, because the license is inside the shrink-wrapped package, there is no way the customer can read it first in order to decide whether or not to accept it, and so the legal enforceability of these "shrink-wrap" licenses is doubtful.

Fair Use

One legal question that should be addressed is whether individual copying of software for personal use might fit within the parameters of the fair use doctrine, which permits the duplication of copyrighted material under some conditions. A determination of fair use must consider four factors: (1) the purpose or character of the use (whether commercial or productive); (2) the nature of the use (whether one is primarily availing oneself of the uncopyrightable factual content of the work or of its expression); (3) the substantiality of the use (whether the work is copied in its entirety or only in part); and (4) the effect of the action on the market for the work. Hornik (1994) argues persuasively that softlifting fails to meet the criteria for fair use. (1) Although the software is not being copied for sale, and even though it may be used in ways that benefit the rest of society (the main aim of fair use), it is also likely to be used indirectly for financial gain. (2) What the softlifter is primarily interested in is the "expression" of the software, that is, its embodiment in code. (3) In softlifting, the entire work is duplicated. Finally, (4) although some softlifters would choose not to buy the software if the choice were to pay or to do without, the practice probably decreases sales at least to some extent. Hence, a fair use defense of softlifting as it is typically practiced would probably not stand up in court.

Software Patents

In many cases, protection of the source code by copyright is insufficient to protect the innovations embodied in the software. These innovations, although representing considerable investment in research and development effort, are often quite easy to reproduce, or "clone." Court cases have upheld the use of copyright protection in some cases in which the "look and feel" of a software product was copied. However, copyright cannot protect against copying of the fundamental ideas and algorithms used in its design. Software makers have therefore turned to patents for this type of protection. Initially, the U.S. Patent Office was reluctant to issue patents for computer software, but a series of court decisions in the 1980s clarified the situation, so that in the 1990s there was a surge in the granting of software patents (Hunt, 2001).

Enforcement Efforts

One area in which the software industry has had some success in improving compliance with licensing and copyright laws has been in pressing its case against corporate softloaders. On behalf of its member companies, SIIA has undertaken a program of voluntary audits of large firms, inspecting their computers to determine whether software has been loaded onto them in violation of license agreements. When SIIA receives a reliable report (usually from an employee of the firm) that softloading is occurring, the firm is sent a letter giving it a choice of permitting the audit or being sued for infringement. Most firms choose the audit in order to avoid the adverse publicity that a lawsuit would entail. They may also have favorable software licensing arrangements that would be at risk if they fail to cooperate. If the audit turns up violations, the firm must destroy the infringing software, purchase replacements for it, and pay a fine. In return, SIIA releases the firm from all further legal claims for prior acts of infringement.

Internet piracy presents greater obstacles to legal enforcement. Suppose someone downloads a software package that has been illicitly placed on a Web site. If the software producer wishes to bring suit to redress this action, it faces a number of practical difficulties (Christensen, 1997). First it must decide whom to sue: the downloader, the intermediary (the service provider hosting the Web site), or the person who uploaded the software in the first place. Next it must decide where to sue: because each of the three parties involved in the action could be located in a different country, the choice of an appropriate venue can be complicated. Finally, assuming the prosecution is successful, it must attempt to recover damages.

Prosecution of an individual downloader is difficult in all of these respects. Web sites usually do not maintain logs of download activity for very long. Even with access to these logs, it can be difficult to identify the downloader, and savvy pirates can masquerade as different users. The downloader may be located in a foreign country that may have weaker intellectual property laws than those in the U.S. Although suit may be brought against foreign nationals in a U.S. court, enforcement of the judgment can be difficult. Finally, the amount of damages that can be assessed for a single count of softlifting may not cover the costs of litigation, and the guilty party may not have the means to pay it anyway. The 1997 U.S. No Electronic Theft (NET) Act addressed this last issue by increasing the civil penalties and for the first time providing for criminal penalties for copyright infringement that is not for financial gain, provided the value of the stolen works is at least $1000. These measures have made it more likely that an action against an individual softlifter could result in significant penalties, including jail time. Recently some high-profile actions have been brought against particularly flagrant softlifters, in hopes of making examples of a few in order to deter others. However, because of the difficulties, as well as for reasons of public relations, software companies have historically tended to avoid legal action against individual softlifters.

A more suitable target for legal action may be the uploader. A single upload can be responsible for thousands of downloads, so the damages that can be assessed can be quite large. Furthermore, if the uploader acted for profit, the penalties are even greater. However, many uploaders post software on warez sites simply for the enjoyment of sharing with others and thumbing their noses at powerful software companies. Identification of the uploader can also be difficult, for the same reasons as with

downloaders. Finally, even if the legal action is successful, an individual uploader may not have the means to pay the penalties, and the software company would have to be satisfied with a moral victory.

Prosecution of the online service provider (OSP) used as an intermediary for software piracy may be a viable option. OSPs operate the computers on which reside bulletin board systems (BBSs), Web pages, newsgroups, and chat rooms, all of which can be used to exchange pirated software. U.S. courts have held OSPs responsible for the copyright infringement activities of their customers under the notions of contributory or vicarious liability. Contributory liability is applicable if the OSP knew about the infringing activity and took no action to prevent it. Vicarious liability can apply if the OSP knowingly made available the means to commit infringements, even if it did not monitor or encourage the infringing activity itself. In response to the threat of legal action, many OSPs now enforce strict policies against infringing activities by their users, although undoubtedly much pirating activity still goes undetected.

Prosecution of the OSP has several practical advantages from the point of view of the software maker: the OSP is an established firm that is easily identified, the amount of damages that can be sought is large, and in the event of a successful suit the OSP is likely to have the resources to pay the judgment.

A recent development that makes prosecution of the intermediary harder is peer-to-peer (P2P) file sharing. The novel idea behind a P2P service is that the files to be downloaded do not reside on a central server, but on the computers belonging to the users of the service. The central server, if there is one, only acts as a go-between by maintaining lists of what files the users have made available for download by others. Once a user has located a particular file on another user's machine, the file is exchanged directly from the one machine to the other without any further involvement on the part of the server. P2P file sharing was pioneered by Napster, which was originally designed to allow users to share music files in MP3 format. However, the P2P protocol can be used equally well to share any sort of digital content and many of the newer P2P services support the sharing of software. Some have also adopted an even more decentralized structure that is much less susceptible to legal action.

Technical Protection Mechanisms

Because legal protections alone have not sufficed, software makers have devised various technical mechanisms to prevent the unauthorized copying of their products. The most commonly used protection mechanisms rely on a special key code that must be entered by the user during the installation process. Typically this key code is provided along with the installation medium in each software package. Key codes do not prevent softloading, because there is nothing to prevent the user from installing the same software on multiple machines. The user can at least be limited to using the software on one machine at a time by means of a key disk or a dongle. A key disk is a special diskette or CD, provided along with the software, that must be inserted into the disk drive during operation of the application program. The program queries the key disk from time to time to continually verify the user's authorization. For the key disk to be effective, of course, it must be difficult to copy by the means at a typical user's disposal. A drawback of key disks is that they prevent the disk drive from being used for other purposes while the application is in use. A related alternative is the dongle, a device that attaches to the parallel, serial, or USB port of the computer. As with a key disk, the application queries the dongle as it runs. Dongles are relatively expensive, typically adding $20 to $30 to the cost of an application, so they are only practical for high-end software.

Media-limited installations are a way to prevent softloading. In these schemes, the installation program counts how many times the application has been installed and refuses to exceed the limit. This method requires the installation medium, or at least a component of it, to be writeable. Also, in order for the protection to be effective, the medium must be difficult to copy by standard means.

Mechanisms such as key disks and media-limited installations were suitable during the 1980s when most software was distributed on floppy disks. As applications grew in size and distribution on CD became the norm, these methods were less appropriate. Also, any copy protection scheme that would prevent legitimate uses such as making archival backup copies or reinstalling the software after a hardware failure irritated customers. The result of the consumer backlash against copy protection was that by the early 1990s, relatively few packaged software applications that were being sold included any protection other than an installation key code.

Protection measures that rely on special hardware, whether key disk, uncopyable medium, or dongle, are not well suited to the present time when much software is distributed via the Internet. Often, the software can be downloaded freely, but contains a "time bomb" that will deactivate it after a trial period such as 30 days. Before that period expires, the customer must register and pay for the software, obtaining a key code that renders the installation permanent.

Unfortunately for the software producers, all of the methods that they have invented to deter the unauthorized use of their products can be "cracked," or circumvented. Copy protection schemes suffice to keep the average user, who has no knowledge of the inner workings of software, honest. It is virtually impossible to devise a scheme that a skilled and dedicated cracker cannot defeat. There is a whole underground society of crackers, individuals who vie to be the first to defeat the copy protection of a newly released program. They are very knowledgeable about computers and programming and are often as skilled as the programmers who produce the software. Some crack programs as a hobby, others do it for profit.

Cracking a program typically involves reverse engineering the binary code, taking it apart to find where the key code is checked or the dongle is interrogated, and bypassing or disabling these sections. If the protection scheme involves cryptography, this only adds to the challenge. Cryptography is the science of scrambling the contents of a file in such a way that it can be unscrambled only by using a secret, randomly chosen key. It is a practical impossibility to crack a well-designed modern cryptographic system by sheer guesswork, even using the fastest available computers. However, a fundamental problem facing

any cryptographic copy-protection method is that the software itself must contain a decryption routine including the key, which a clever cracker can in principle discover no matter how well it is hidden.

Recent Developments

The battle against online piracy is no longer the province of the software producers alone. The entertainment media industry is getting involved, due to the increasingly digital nature of their products. Under strong lobbying from Hollywood, proposed new legislation has been introduced in the U.S. Congress that would require copy-protection mechanisms to be embedded in every digital device and in all software that will be produced in the future. The new measures being advocated go even farther than the software industry wishes. Some of the proposed protection mechanisms will interfere with the legitimate duplication of software by OEMs and may prevent computer users from performing legitimate tasks. It is not clear at this time where these efforts will lead, but it is unlikely that they will be any more successful than previous measures in completely stopping piracy.

CONCLUSION

Piracy costs the software industry billions of dollars annually in lost revenues. The precise cost cannot be ascertained, because there are many economic factors that would change if the illicit copying could be stopped entirely. Softlifting and softloading probably account for the largest proportion of the activity, and the Internet is an increasingly important medium for exchange of "warez."

Technical means of enforcing copy protection can always be defeated. Veterans of the struggle against crackers recognize that, at best, copy protection will only slow pirates down and put some obstacles in the softlifters' path, so that enough people will purchase the product for it to be profitable. Furthermore, copy protection, if it is too intrusive, annoys customers and can even backfire by spurring more circumvention efforts. Consequently, most software producers have decided not to rely solely on technical means but to undertake a campaign of user education, coupled with high-profile legal action, to try to persuade customers to obey the laws protecting software. These efforts have borne fruit, reflected in a slow but steady decline in piracy rates in most countries. The biggest reductions in piracy rates have occurred in countries that have been making the transition to free-market economies and developing their own indigenous software industries. Although software piracy will never be completely eliminated, there are good reasons to hope that in coming years it will decline to levels that the software industry can live with.

GLOSSARY

Crack As a verb, to circumvent technical measures intended to prevent the unlicensed operation of a program. As a noun, a program that has been cracked so that it can be used by unauthorized users.

Dongle A specialized hardware device that attaches to a computer's parallel, serial, or USB port and that is queried by a program during operation in order to verify the user's authorization to use the program.

Hard-disk loading The installation of unauthorized software on computers being prepared for sale by original-equipment manufacturers or other computer vendors.

Mischanneling The selling of software intended for academic, government, or other special categories of customers to those who do not belong to the intended group.

Softlifting The unauthorized copying of software by an end user for his or her own use, rather than for sale.

Softloading The copying of software by an end user onto more machines than permitted by the license, or the unauthorized loading of software onto a server for use by client machines in a local-area network.

Unbundling The selling of software that is licensed only to be sold as part of a package as a separate item.

Warez Slang term for pirated software, usually referring to items made available on the Internet.

CROSS REFERENCES

See *Copyright Law; Legal, Social and Ethical Issues; Patent Law; Trademark Law.*

REFERENCES

Business Software Alliance (1999). *1999 global software piracy report*. Retrieved July 11, 2002, from http://www.bsa.org/usa/globallib/piracy/1999_Piracy_Stats.pdf

Business Software Alliance (2002). *Seventh annual BSA global software piracy study*. Retrieved July 11, 2002, from http://www.bsa.org/resources/2002-06-10.130.pdf

Cheng, H. K., Sims, R. R., & Teegen, H. (1997). To purchase or to pirate software: An empirical study. *Journal of Management Information Systems, 13*(4), 49–60.

Christensen, K. D. (1997). Fighting software piracy in cyberspace: Legal and technological solutions. *Law & Policy in International Business, 28*, 435–475.

Davis, R., Samuelson, P., Kapor, M., & Reichman, J. (1996). A new view of intellectual property and software. *Communications of the ACM, 39*, 21–30.

Hornik, D. M. (1994). Combating software piracy: The softlifting problem. *Harvard Journal of Law & Technology, 7*, 377–417.

Hunt, R. M. (2001, First Quarter). You can patent that? *Business Review,* 5–15.

Husted, B. W. (2000). The impact of national culture on software piracy. *Journal of Business Ethics, 26*, 197–211.

Marron, D. B., & Steel, D. G. (2000). Which countries protect intellectual property? The case of software piracy. *Economic Inquiry, 38*, 159–174.

PricewaterhouseCoopers (1999). *Contributions of the packaged software industry to the global economy*. Retrieved April 5, 2002, from http://www.bsa.org/usa/globallib/econ/pwc1999.pdf

Simpson, P. M., Banerjee, D., & Simpson, C. L. Jr. (1994).

Softlifting: A model of motivating factors. *Journal of Business Ethics, 13*, 431–438.

Slive, J., & Bernhardt, D. (1998). Pirated for profit. *Canadian Journal of Economics, 31*, 886–899.

Software and Information Industry Association (2000). *SIIA's report on global software piracy 2000.* Retrieved December 21, 2001, from http://www.siia.net/piracy/pubs/piracy2000.pdf

Taylor, G. S., and Shim, J. P. (1993). A comparative examination of attitudes toward software piracy among business professors and executives. *Human Relations, 46*, 419–433.

Speech and Audio Compression

Peter Kroon, *Agere Systems*

INTRODUCTION

Audible signals such as speech and music are acoustic analog waveforms, pressure changes that propagate through a medium such as air or water. The waveforms are created by a vibrating source such as a loudspeaker or musical instrument and detected by a receptor such as a microphone diaphragm or eardrum. An example of a simple waveform is a pure tone, a periodic signal that repeats many times per second. The number of repetitions per second is its *frequency* and is measured in Hertz (Hz). Audible tones are typically in a range from 20 to 20,000 Hz, which is referred to as the *bandwidth* of audible signals. The tone will create a sound pressure displacement that is related to its amplitude. Signals with high amplitude will sound louder than signals with low amplitude, and the range from soft to loud is called the *dynamic range*. Complex sounds (e.g., the sound of a piano, or speech) consist of combinations of many tones of different frequencies and amplitudes that vary over time.

Using a microphone one can capture an acoustic waveform and convert it into an electric signal or waveform. This signal can be converted back to an acoustic signal by using a loudspeaker. To represent this analog waveform as a digital signal, it is necessary to find a numerical representation that preserves its characteristics. The process of converting an analog signal to a digital signal is usually referred to as *digitization*. Digital representation of audio and speech signals has many advantages. It is easier to combine with other media such as video and text, and it is easier to make the information secure by applying encryption. Digital representations also allow procedures to protect against impairments when transmitting the signals over error-prone communication links. The main disadvantage is that straightforward digitization of analog signals results in data rates that require much more capacity of the physical channel than the original analog signal.

Before we provide some examples of this dilemma, let us first take a look at the principles of digitization. To digitize an analog audio signal it is necessary to sample the signal at discrete instants of time at a rate equivalent to twice the highest bandwidth that exists in the signal (this is the *sampling* or *Nyquist theorem*). The frequency that the signal is sampled with is referred to as the *sampling frequency*. Typical sampling frequencies for speech signals are between 8 and 16 kHz, whereas for music signals ranges between 16 and 48 kHz are more common. To get a digital representation, the sample values need to be a discrete set of numbers represented by a binary code. This process is referred to as *quantization*. In contrast to sampling, which allows one to perfectly reconstruct the original analog waveform, quantization will introduce errors that will remain after the analog signal is reconstructed. The quantization error (defined as the difference between the analog sample value and the discrete value) can be made smaller by using more bits per sample. For example, an 8-bit number allows 2 to the power of $8 = 256$ different values, whereas a16-bit number allows 65,536 different values. For speech signals between 8 and 16 bits per sample are adequate, whereas for high-quality music signals between 16 and 24 bits per sample are commonly used.

The process of sampling and quantization described above is referred to as pulse coded modulation (PCM). The total bit rate per second for a PCM signal is given by (sampling rate) $\times$ (number of bits per sample) $\times$ (number of audio channels).

For a stereo signal on a compact disc this means $44{,}100 \times 16 \times 2 = 1{,}411{,}200$ bits per second (1,411 Mb/s). As is illustrated in Figure 1, the typical bit rates needed for various signals can be quite high. For storage and transmission purposes these rates quickly become prohibitive. Although disc-based storage has become cheaper and high-speed Internet connections are more commonplace, it is still difficult to stream compact disc data directly at about 1.4 Mb/s or to store hundreds of uncompressed CD's on a hard disk.

The main goal of *audio and speech compression* is to find a more efficient digital representation of these signals. Because most signals start as simple sampled PCM signals, it is useful to use the resulting relative reduction in bit rate as a measure of efficiency. It should be pointed out that a reduction in bit rate is not always the main objective. For example, one could increase the bit rate to make the signal more robust against transmission errors. In that case the generic term *coding* is more appropriate. In this chapter we will use both terms interchangeably. Figure 2 shows the generic compression operation. The *encoder* takes the PCM input signal and generates the compressed *bit stream*. The bit stream is either transmitted to the

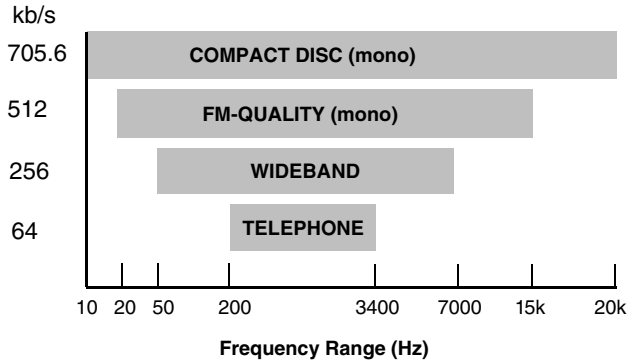


Figure 1: Relationship between bandwidth and bit rate.

decoder or stored for later retrieval. The decoder takes the bit stream and generates the corresponding decoded PCM signal, which is a rendering of the original input PCM signal.

In the remainder of this chapter we will focus on reducing the bit rates of speech and audio signals while providing the best possible signal quality. At this point it is good to point out that quality is a difficult attribute to quantify because it has many dimensions. Most of these are associated with specific applications, and it is important to set proper objectives when designing or choosing a particular compression algorithm. For example, for speech communications it could be important to have consistent performance for various input conditions, such as clean speech, noisy speech, and input level variations. For compression of music signals it could be important to have consistent performance for various types of music, or to preserve audio bandwidth and stereo image as much as possible. As will be clear later, this quality objective will be constrained by other factors such as *delay* (the time needed to encode and decode a signal) and the *complexity* (the number of arithmetic operations) of the methods used.

Speech and audio compression applications can be divided into two classes. The first is *broadcasting* (e.g., streaming); the other is *communication* (e.g., Internet telephony). Each application has different requirements, as can be seen from Table 1.

There are two principal approaches to the compression of digital signals: *lossless* and *lossy* compression. Lossless compression techniques take advantage of redundancies in the numerical representation. For example, instead of using 16 bits per sample uniformly, one could use a new mapping that assigns symbols of shorter length to the most frequent values. Or if the signal values change slowly between sample values, one could encode the differences instead, using fewer bits. Lossless compression is a reversible operation and the input and output signal samples of Figure 2 will be identical. Lossy compression techniques, on the other hand, assume that the signal has

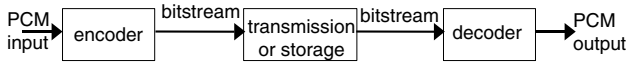


Figure 2: Block diagram of a generic coding or compression operation.

Table 1 Difference in Requirements for Broadcasting and Communications

Application	Broadcasting	Communications
Characteristic	One-way transmission	Two-way transmission
Delay	Not important	Important
Complexity	Not important	Important
Technology	Audio coding	Speech coding

a human destination, which means that signal distortions can be introduced as long as the listener either is not able to hear them, or has no serious objections. Lossy compression is not reversible and the input and output signal samples of Figure 2 will be different. For many signals this difference would be unacceptable, but for audio signals, one only worries about audible differences. If the differences are inaudible the lossy coding techniques used are often referred to as *perceptually lossless* coding techniques. But even if the differences were clearly audible it still would be acceptable for many applications. A good example is the difference between telephone speech and natural speech, where the telephone signal is significantly limited in bandwidth (typically less than 4 kHz).

In practice, the use of lossless coding for audio and speech signals results in limited compression efficiency and its use is restricted to high-quality applications such as the audio DVD. The compression efficiency for lossy coding can be significantly higher and consequently perceptually lossless and lossy coding are the main approaches used in most audio and speech compression applications. Speech compression takes this approach one step further by also assuming that humans generate the source signal, which gives it certain properties that can be taken advantage of by the compression algorithms. As a result, speech compression can achieve very high compression efficiency. Figure 3 gives a summary of the typical bit rates and applications.

Compression for Packet Networks

The compression algorithms described in this chapter are used for communication and broadcasting over wired and

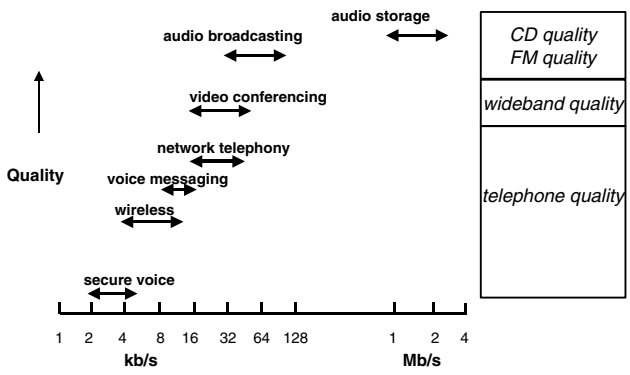


Figure 3: Typical bit rates and applications for speech and audio compression.

wireless networks. Errors that get introduced into the bit stream during transmission can introduce serious degradations in the decoded signal. In contrast to analog signals, where transmission impairments mainly introduce additional noise into the audio signal, digital signals subjected to bit stream errors will produce pops, clicks, and other annoying artifacts. Especially in wireless applications, one is likely to encounter transmissions errors, and it is common to add additional information for *error correction and detection*. However, even with the use of error correction techniques, it is still possible that bit errors remain in the bit stream.

The sensitivity of a decoder to random bit errors (in other words, their relative impact on the decoded signal quality) should be taken into account in such an application. For wired networks the transmissions channels are usually good, and transmission errors are unlikely. However, in packet networks (e.g., the Internet), it is possible that packets will not arrive on time. Due to the real-time nature of the connection, the decoder cannot request retransmission and this information is considered to be lost. To avoids gaps in the signal it is necessary to fill in the missing information. This technique is referred to as *error mitigation*. Sophisticated error mitigation techniques work quite well for segments up to 40–50 ms. For longer error bursts it is necessary to mute the signal. Is it important to realize that for applications where transmission errors can occur, the overall quality of a coder (including the use of error correction and mitigation) may be dominated by its robustness to channel impairments.

Traditional compression applications are optimized for the underlying application (e.g., a cellular system). These systems are homogeneous, in the sense that all terminals and links meet certain minimum requirements in throughput and capabilities. The Internet is a much more heterogeneous network, where endpoints can be quite different in capabilities (e.g., low-end vs. high-end, PC vs. laptop, wired vs. wireless) and connection throughput (dial-up vs. broadband). One solution would be to use scalable coders in which the same coding structure can be used for operation at different bit rates. This requires a handshaking process between transmitter and receiver to agree on the rate to be used. Moreover, if a throughput issue came up somewhere in the middle of a link, it would require decoding at the higher rate first, and then subsequent encoding at the lower rate. The resulting *transcoding* operation introduces additional delay and a significant degradation in quality, because coding distortions are compounded. A better approach is the use of *embedded* coders. In embedded coders there is a core bit stream that each decoder needs to decode the signal with a certain basic quality. One or more enhancement layers enhance this core layer. Each enhancement layer will increase the average bit rate and the quality of the decoded signal. The encoder generates the core layer and enhancement layers, but the decoder can decode any version of the signal as long as it contains the core layer.

This is illustrated in Figure 4. Apart from adjusting for various throughput rates, embedded coders can also be used for temporary alleviation of congestion. If too many packets arrive at a given switch (e.g., switch 2 in Figure 4), the switch can decide to temporarily drop some enhancement packets to avoid congestion. Depending on the size of the enhancement layer bit streams and the coder design, this can be done with only a minor impact on the audio quality. The embedded coding approach will alleviate packet loss problems but will not avoid them. Because we always need the core information, the packet loss problem will remain. A solution to this problem is the use of multidescriptive coders. Multidescriptive coders can be seen as a superset of embedded coders. In this the encoder creates two or more descriptions of the signal, each of which can be decoded independently. When all descriptions are received the best possible quality is received; when only one of the descriptions is received the quality will be lower. In practice, it is difficult to design efficient scalable coders and multidescriptive coders, and this topic is still an active area of research.

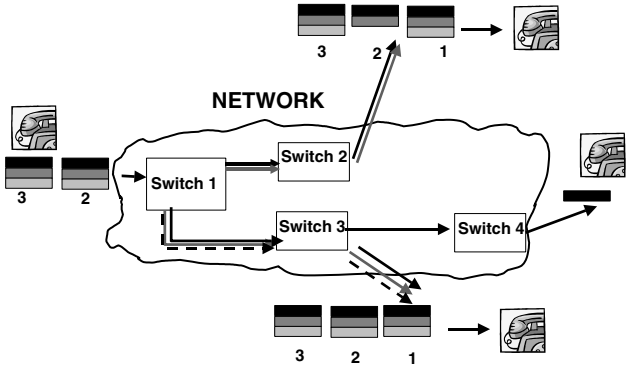


Figure 4: Use of embedded coders in a heterogeneous network.

Speech and Audio Quality Assessment

The quality assessment of lossy speech and audio compression algorithms is a complicated issue. Quite often we want to assess not only the quality of the compression algorithms, but also the quality delivered by these algorithms in a typical operating scenario, which means including other aspects of the delivery chain as well, such as the quality of the network, or the quality and type of playback equipment. Because the coders use perceptual techniques, it is important to use human listeners. Even in this case care has to be taken to get reliable and reproducible results. Choices of test material, number of listeners, training of listeners, test format (e.g., order of playback, inclusion of original), and playback scenario (e.g., headphones vs. loudspeakers) all affect the outcome of the test, and it is necessary to design the test so that the impact of these factors can be minimized. Testing perceptually lossless coders (no audible differences when compared to original) will be different from assessing the performance of lossy coders. The latter will be complicated because tradeoffs have been made in many dimensions, which will produce different listener responses. For example, some people prefer large audio bandwidth to reduced signal distortion. Some stereo audio coders trade off more subtle issues such as stability of the reproduced stereo image. A well-designed test will try to eliminate all these biases and should produce a reproducible result.

Assessment of speech is even more involved, because it involves two-way communication and most material is context-sensitive. Mean opinion scores (MOS) testing in which panels of listeners rate the quality of short sentences on a 5-point scale is the most common. Besides speech quality one can also test speech intelligibility, although this is usually not an issue for telephony bandwidth speech at bit rates of 4 kb/s and higher.

It should be appreciated that none of the testing methods described above can fully predict how people will experience the quality in a real-world scenario, which involves talking to people whose voices they know or listening to music they like.

Tests with human subjects are expensive and time-consuming, and one would like to use objective measures that could predict subjective quality based on a comparison between the original and the processed version. For lossy compression simple objective measures such as segmental signal-to-noise (SNR) measurements are meaningless. A more effective approach is to include models that mimic our auditory system and use the resulting model output to predict subjective quality. Two standards based on such an approach have recently been recommended: ITU-R PEAQ for the assessment of audio compression techniques and ITU-T P.862 for the assessment of telephony-quality speech compression techniques. Although these methods have shown to be quite accurate for some scenarios, they should always be used with caution and with a clear understanding of their shortcomings.

SPEECH CODING TECHNIQUES

Efficient digital presentations of speech signals have been a topic of research since the 1940s, but only since the early 1990s have many applications become technically and economically feasible. Digital cellular telephony has been one of the main applications for speech coding, and many digitization choices were made to be compatible with wired digital networks. For example, the speech signals are sampled at 8 kHz (thereby limiting the signal bandwidth to 4 kHz), single channel (mono), and 8 to 16 bits/sample. The communication application puts a constraint on the *delay* introduced by the compression operation. Not only is it is difficult to have a natural two-way conversation with delays exceeding 250 ms, but it is also more noticeable to hear echoes introduced, for example, by the acoustic coupling between loudspeaker and microphone (e.g., a speakerphone) at either end of the communication link. For conferencing applications, which involve more than two parties, each party will hear the combination signal for all other participants. Because this combining of the signals needs to be done in the PCM domain it is necessary to decompress the signals, digitally combine them, and compress them again. The delay introduced by compression will now be compounded, thereby reinforcing the problems mentioned above. Most compression algorithms introduce delay because they analyze the signal in blocks or frames with a duration of 10 to 30 ms. Analysis in frames is necessary to better characterize the signal behavior and its variations.

For communication applications it is also important to put constraints on the *complexity* of the compression operation because each end-point needs both an encoder and a decoder. This is even more relevant for wireless applications where the end-point is battery-powered and high complexity will reduce battery life. Complexity is defined in terms of computational load (MIPS) and memory usage (RAM and ROM). For most speech and audio coding algorithms there is an asymmetry in complexity, and the encoder can be several more times complex than the decoder.

Most speech coders are based on the lossy compression paradigm, taking advantage of the properties of the auditory system and the properties of the speech production mechanism. The latter can be taken advantage of by using so-called *parametric coders*. With these coders the speech signal is modeled by a limited number of parameters, which are usually related to the physical speech production mechanism. The parameters are obtained by analyzing the speech signal and quantized before transmission. The decoder will use these parameters to reconstruct a rendering of the original signal. When the input and output waveforms are compared, the resemblance may be weak but the signals may sound very similar. Using parametric approaches, it is feasible to achieve reasonable quality with very low bit rates (2 to 4 kb/s). The quality is limited by the accuracy of the model. This is illustrated in Figure 5. To avoid this limit on quality, a more common approach is to use waveform-approximating coders. These coders maintain the waveform of the original signal as much as possible while taking advantage of the properties of both the speech production and auditory mechanisms. The resulting quality is better at the expense of higher bit rates, and at lower bit rates the quality of a waveform coder will be less than that of a parametric coder operating at the same low rate (see Figure 5).

Speech as produced by humans has certain properties that can be taken advantage of for compression. It has limited energy above 8 kHz, and it has a limited dynamic range. This allows sampling with frequencies between 8 and 16 kHz and PCM quantization with 12–16 bits/sample. Using nonuniform quantization (e.g., the quantizer step sizes are small for small input values and large for large input values), it is possible to quantize telephone speech with 8 bits per sample. Figure 6 shows a waveform and its corresponding spectrogram. From looking at this figure one can see that the envelope of the amplitudes change slowly as a function of time. The spectrogram shows that certain frequencies are stronger than others. These emphasized frequencies are called *formants*

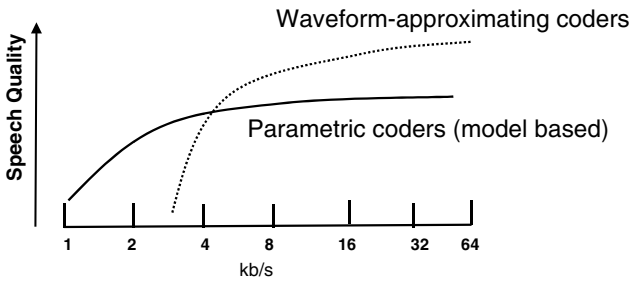


Figure 5: Quality vs. bit rate curves for waveform and parametric coders.

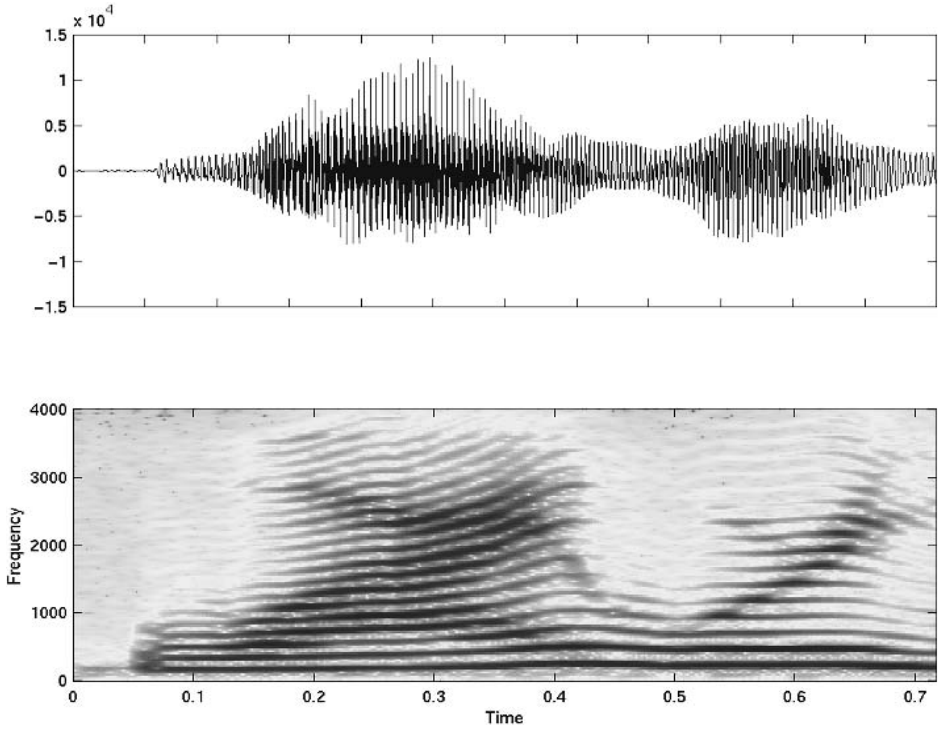


Figure 6: Time waveform of the utterance "Why were you away a year, Roy?" spoken by a female talker (top) and corresponding spectrogram (bottom). The spectrogram shows the power spectrum as a function of time, where the gray level indicates the power level, from low (white) to high (black).

and are caused by resonant frequencies in the vocal tract in response to the glottal excitation signal. Changing the shape of the mouth cavity over time changes the resonant frequencies. The relative position of these formants defines the consonants and vowels that we perceive. One can also see a more harmonic component, which is due to the periodic excitation provided by the vocal cords. This is referred to as the *pitch* of a sound and varies between 100 and 250 Hz for males and 200 and 400 Hz for females.

All these features are produced by the human articulatory system, which changes slowly in time, because it is a biological system driven by slowly moving muscles. As a result it was found very effective to represent the formants by a slowly varying adaptive digital filter, the so-called *linear or short-term prediction* filter. With only 10 predictor coefficients updated once every 20 ms it is possible to accurately reproduce the evolution of speech formants.

Figure 7 shows a short segment of a speech signal and the corresponding signal after filtering with an adaptive linear prediction filter, called the *residual* signal. Because this signal is better behaved it usually easier to quantize with fewer bits per sample. The predictor coefficients have to be quantized before transmission, and numerous methods are available. The most effective methods require only 1,500 to 2,000 bits/s to transmit 10 coefficients every 10 to 20 ms. By recognizing the periodicity in the signal for voiced sounds (e.g., vowels) it is possible to further improve predictor efficiency. A so-called *long-term or pitch predictor* is able to predict the periodic component, resulting in an even more noiselike residual signal. The long-term predictor consists of a variable delay line with a few filter coefficients. The delay and coefficients are updated once every 5 to 10 ms. The most typical configuration is forward adaptive, which means that about 1,500 to 2,000 bits per second are needed for transmitting the long-term predictor parameters. Figure 8 shows the signals after short-term and long-term prediction, respectively.

The signal shown at the bottom of Figure 8 resembles a noiselike signal, with a reduced correlation and reduced dynamic range. As a result it can be quantized more efficiently.

At this point we have all the components needed for a linear predictive coder. A block diagram is shown in Figure 9. The signal is filtered with the short-term filter $A(z)$ and long-term filter $P(z)$ and the remaining residual signal is quantized. The predictor parameters and quantized residual signal are transmitted or stored. The decoder, after decoding the quantized residual signal, filters it through the inverse long-term and short-term prediction filters. Note that without quantization of the residual signal the decoder can exactly reproduce the original signal. For quantization of the residual signal many techniques exist. The simplest quantizers are scalar uniform or nonuniform PCM quantizers. For acceptable results at least 4 to 5 bits per sample are needed. Even refinements such as adaptive quantization, in which the quantizer step sizes are adjusted over time, will not reduce the number of bits per sample significantly. More efficient quantization can be obtained through the use of *vector quantization* (VQ), in which multiple samples are quantized simultaneously.

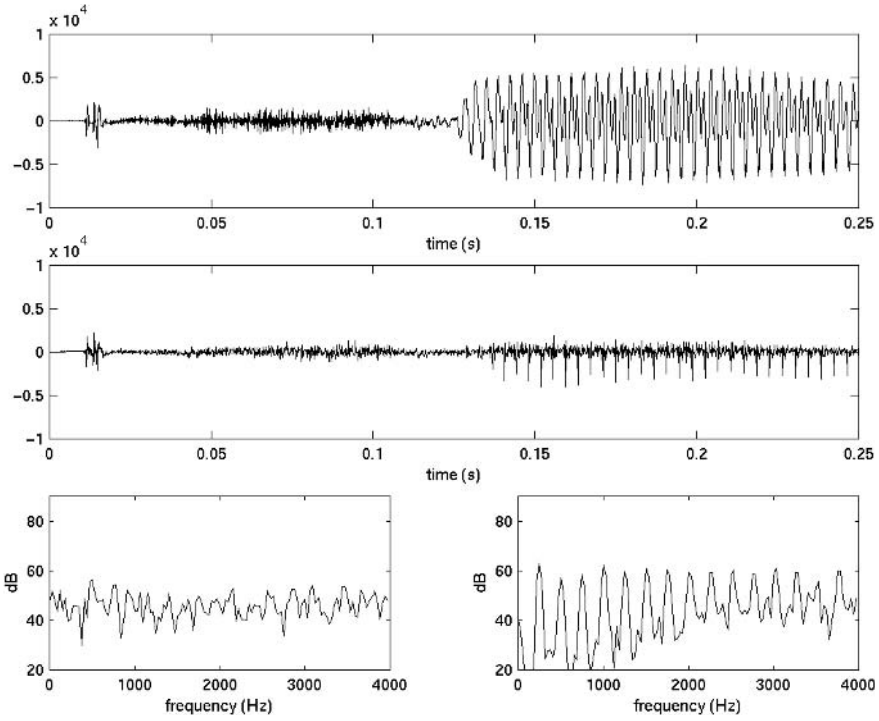


Figure 7: Time waveform (top), corresponding residual signal (middle), and spectra of the residual signal (bottom) for the first part of the sentence (left) and second part of the sentence (right).

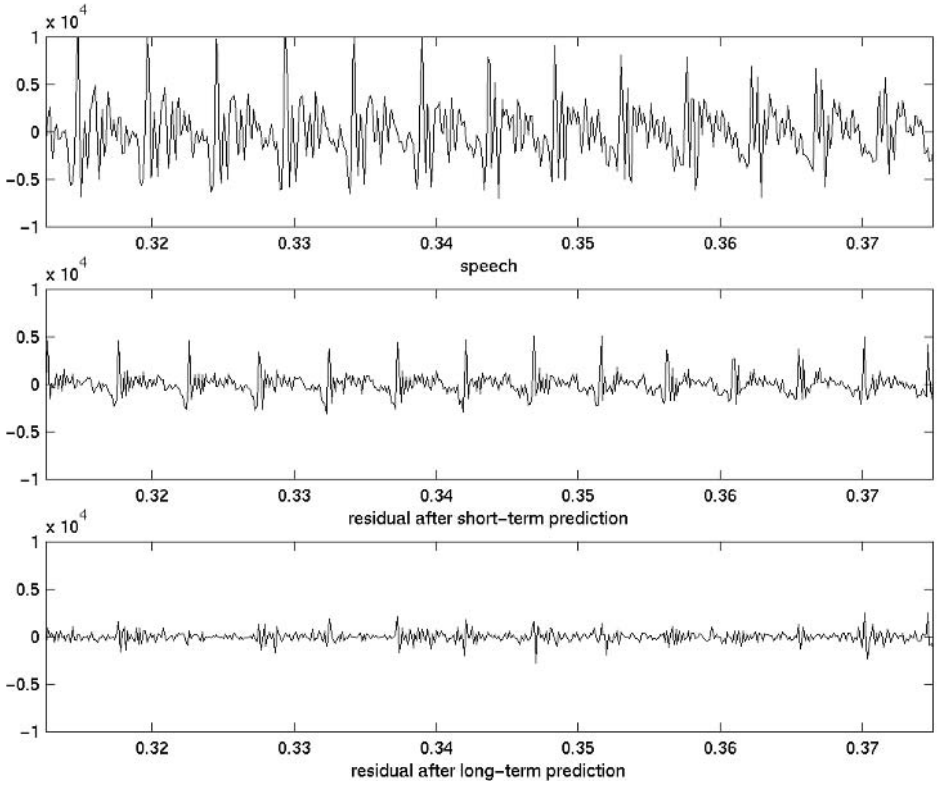


Figure 8: Time waveform (top), short-term residual signal (middle), and long-term residual signal (bottom).

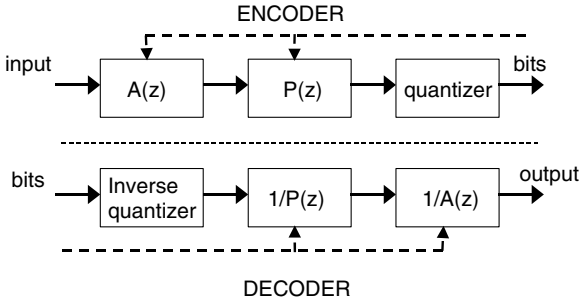


Figure 9: Block diagram of a linear predictive coder with short-term and long-term prediction.

This is achieved by having a number of possible representative sample sets or *codebooks*. Quantization is achieved by selecting the codebook entry (each entry containing multiple samples) that is the best representation of the signal to be quantized. The codebook index corresponding to this entry is transmitted and the decoder, which has a similar codebook, uses this index to look up the corresponding values. Whereas the lowest quantization rate for scalar quantization is 1 bit per sample, VQ allows fractional bit rates. For example, quantizing 2 samples simultaneously using a 1-bit codebook will result in 0.5 bits per sample. More typical values are a 10-bit codebook with codebook vectors of dimension 40, resulting in 0.25 bits/sample. Both scalar and vector quantization attempt to find the quantized value that is the closest to the original unquantized input. In the block diagram of Figure 9 one can argue that the overall goal is not to find the best quantized residual values, but the best quantized speech signal. Especially for coarse quantization (very few bits per sample), this becomes an issue. A powerful technique used in speech coding is *analysis-by-synthesis*, in which the effect of quantization is determined by examining the effect on the decoded output. This is accomplished by operating the diagram of Figure 9 in the configuration shown in Figure 10 .

In this figure the encoder of Figure 9 is enhanced with a local decoder. In most practical coders the predictors are still computed as usual. The quantization of the residual signal is done in an analysis-by-synthesis fashion. If we assume that we use a codebook, then the essence of this approach requires that for each codebook vector we perform local decoding and compare the resulting prototype output with the original input signal. The codebook vector that gives the best approximation is selected. Note that with this paradigm we are indirectly creating a quantized residual signal, and this signal is often referred

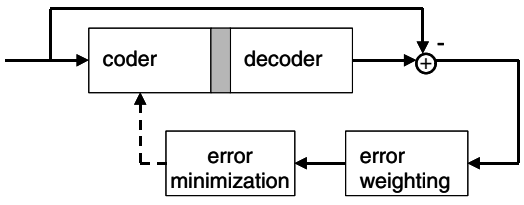


Figure 10: Analysis-by-synthesis encoder with error weighting.

to as an *excitation* signal. Instead of directly comparing the original input signal and the quantized and decoded rendering of this signal, an error-weighting filter is introduced, which better reflects the way our auditory system perceives distortions. It should be noted that this weighting is much simpler than the complex auditory models used in audio coding. The block diagram of Figure 10 forms the basis for a family of coders generically referred to as code-excited linear predictive (CELP) coders. Most modern speech-coding standards are based on this principle. A large amount of research has been done on efficient codebook structures, which not only give the best performance, but are also manageable in terms of search. A commonly used structure is the so-called *algebraic* codebook, which consists of a few nonzero pulses with deterministic positions.

For all coders several techniques can be used to further improve their performance. Some of these techniques can be done independent of the coder, although in practice it makes sense to take advantage of the parameters already computed by the coder. Preprocessing techniques that are useful are gain control and noise suppression. The latter can be quite sophisticated but very effective, especially for the lower rate speech coders, which typically do not handle background noise well. A widely used form of postprocessing is *postfiltering*. In this process the decoded speech signal is slightly distorted in such a way that the coding noise gets suppressed and the signal gets enhanced. If done with care it can clean up a signal, resulting in perceived quality improvement.

Another technique that has found some popularity is taking advantage of the fact that conversational speech comes in bursts, due to the speakers talking at alternate times. Sometimes there can be large pauses in between words. This can be taken advantage of by only transmitting when active speech is present. When speech is not active no signal is transmitted. Because on the average people speak half of the time, this technique has the potential to reduce the bit rate by half. To make this *discontinuous transmission* approach work, a *voice activity detector* (VAD) is needed. For speech without background noise, this approach can work quite well. When background noise is present (e.g., a car), it is more difficult to get reliable decisions from the VAD. Moreover, when no talker is active, and no signal is transmitted, the receiver side needs to substitute a replacement signal. This is referred to as *comfort noise*. For high levels of background noise it is difficult to have this comfort noise match its characteristics. Hence more sophisticated systems transmit low-rate information about the background noise, such as energy and spectral characteristics, at average rates of about 1 to 2 kb/s.

Speech Coding Standards

For communication purposes it is important to establish standards to guarantee interoperability between equipment from different vendors, or between telecommunication services in different geographic areas. Telecommunication standards are set by different standard bodies, which typically govern different fields of use.

Table 2 Summary of Relevant ITU-T Coding Standards: Quality and Complexity Are Indicated by Asterisks, Where a Single Asterisk Means Low, and More Asterisks Mean Increasing Complexity

Coder	Algorithm	Rate Kb/s	Frame size	Complexity	Quality
G.711	Mu/A-Law	64	1	*	*****
G.726	ADPCM	32,40,16,24	1	**	*****
G.727	ADPCM	32,40,16,24	1	**	*****
G.728	LD-CELP	16	5	*****	*****
G.729	CS-ACELP	8	80	****	****
G.723.1	MP-MLQ	6.3	240	****	***
	ACELP	5.3	240	****	**

The International Telecommunications Union, ITU-T, establishes worldwide telephony and communication standards, whereas regional standards bodies such as ETSI and TIA define standards that are more regional in character (e.g., wireless standards). For Internet applications there is a proliferation of ITU-T standards and protocols. Table 2 summarizes the most relevant ITU-T speech coding standards.

G.729 is one of the more commonly used coders in VoIP applications. Besides the main standard there are many extensions to this standard, which in ITU-T terms are referred to as annexes. Annex A, for example, is a low-complexity version of G.729, which generates bit-compatible output. It requires only about half the complexity, at the expense of minor degradation in speech quality. Other annexes of G.729 define a low-bit-rate version and a high-bit-rate version and integration with Voice Activity Detection.

Most of the discussion has focused on coders for an 8-kHz sampling rate, which limits the audio bandwidth to 4 kHz. With the availability of new endpoints, it is now more feasible to provide higher quality output, specifically increased audio bandwidths. A commonly used sampling rate is 16 kHz, which supports audio bandwidths up to 8 kHz. Table 3 summarizes several ITU-T standards that have been defined for encoding these so-called wide band signals.

Most of these ITU-T standards not only come with detailed technical descriptions of their underlying algorithms but also are accompanied by reference source code and test vectors. The source code is provided as floating point C code or fixed point C code or both. Fixed-point code reflects implementation on signal processor and VLSI chips, which are typically used in most portable devices.

Although standards make the technology readily available, they are not license-free. In most cases proper royalty agreements must have been obtained before one can use the coder in a commercial application.

AUDIO CODING TECHNIQUES

The most common high-quality audio distribution format is based on the compact disc format introduced in the early 1980s. The signal is encoded using PCM with 16 bits/sample using a 44.1-kHz sampling rate. For stereo signals this means a data rate of $44{,}100 \times 16 \times 2 = 1.41$ Mb/s. More recent formats such as DVD-audio support up to 24 bits/sample, multichannel formats (e.g., 5.1), and sampling rates up to 192 kHz, resulting in even higher data rates. For most practical purposes these signals will be used as digitized source signals. For Internet streaming and computer storage applications, it is necessary to reduce these rates significantly and to bring them into the 32 to 128 kb/s range. As discussed earlier, this can be accomplished by the use of perceptually lossless coders, which take advantage of the limitations of our auditory system. For CD quality it is possible to make signals sound perceptually indistinguishable from the original at 64 kb/s per channel (128 kb/s for stereo). At lower rates we lose some of the information, but if this is done by proper combinations of bandwidth reduction, reduced dynamic range, and the use of mono instead of stereo, the resulting signal will still be acceptable for many applications. Perceptually lossy and lossless compression uses two main techniques. First we have *irrelevancy* removal, which removes parts of the signal that we cannot hear. The second technique, *redundancy* removal, finds the most compact signal representation.

Irrelevancy removal exploits the properties of the human auditory system. The human auditory system is a highly sophisticated system with tremendous capabilities. It acts as a converter of acoustic waves to auditory nerve firings, while performing a spectral analysis as part

Table 3 ITU-T Wideband Coders for 16-kHz Sampling Rate

Coder	Algorithm	Rate Kb/s	Frame size	Complexity	Quality
G.722	ADPCM	64,56,48	1	*	*****
G.722.1	Transform	32,24	320	***	****
G.722.2	CELP	15.85, 6.6–23.05	320	****	***** – ****

of this process. The auditory system has been shown to have masking properties. Masking describes the process in which one signal becomes inaudible in the presence of another signal. In other words, under certain circumstances it is possible to make quantization noise inaudible while the decoded audio signal is present. Masking can happen both in time and in frequency. Understanding the principles of masking has taken many decades of research using both physiological and physical measurements. Masking data are obtained through psychoacoustical studies, in which subjects are exposed to test signals and asked about their ability to detect changes (increase in frequency, audibility, etc). Most of the understanding of masking is based on simple tones and noise. Because complex signals can be viewed as composites of time-varying tones, a common approach has been to derive the masked threshold by analyzing the signal tone by tone in specific frequency bands, related to the frequency bands used by the human auditory system to analyze sounds. These bands, called critical bands, are spaced nonuniformly with increasing bandwidth for higher frequencies. In each critical band, the signal and its corresponding masking function are calculated, and the masked threshold is derived as a superposition over the complete frequency band. It should be noted that the actual procedure is much more complicated, taking into account several interactions and characteristics of the signals (e.g., if the signal is noiselike or tonelike). Figure 11 gives an example of the power spectrum of a signal and its corresponding masked threshold. In this figure, as long as the quantization noise remains below the solid line, it will be inaudible.

In general the model that is used to derive the masked thresholds is referred to as the *psychoacoustic* or *perceptual model*. Building a good perceptual model and matching it properly to a given coder structure is a very complex task. It also should be noted that for a given coder structure (or coder standard such as MPEG 1, Layer 3), it is possible to improve the perceptual model while still being compliant to the standard. This also explains the quality differences in various encoders that all support the same standard.

To achieve redundancy removal, and at the same time take advantage of the frequency-domain masking properties, it is beneficial to perform a spectral decomposition of the signal by means of a filter bank or transform. Most modern audio coders are based on some form of lapped transform, which not only provides computational efficiency, but also allows perfect reconstruction. In other words the transform and its inverse will produce a delayed version of the original time signal. The transform sizes and overlaps will be chosen in such a way that the signal is critically sampled (i.e., the number of frequency components is the same as the number of time samples). The size of the transform will determine the spectral resolution. A larger transform will provide better spectral resolution at the expense of decreased time resolution. A common solution is to make the transform size adapt to the signal. For stationary signals, a large transform size is chosen, whereas for nonstationary signals or onset's a smaller transform size is chosen. Typically sizes vary from 256 to 2,048 samples for sampling rates in the range 20 to 40 kHz. This process is called window switching, and care is taken to make sure that the whole process is invertible. A commonly used transform is the modified discrete cosine transform (MDCT). It uses a transform of length 2M samples, which advances M samples between adjacent transforms. It is critically sampled and only M coefficients are generated for each 2M set of input samples. Computationally efficient implementations have contributed to the widespread use of the MDCT in many audio coding standards.

Figure 12 shows a generic block diagram of an audio encoder incorporating the filter bank and the perceptual model. The resulting spectral components (MDCT coefficients) are quantized in such a way that the resulting quantization noise is inaudible. This is accomplished by using the masking level obtained from the *perceptual model*. The amount of noise is controlled by the resolution of the quantizer. By choosing a different quantizer step size the amount of noise can be quantized. Typically the same quantizer is used for a set of coefficients, and the corresponding step size for that set is transmitted to the decoder. It should be noted that at this point, due to quantization, the decoded signal will be different from the original (i.e., lossy coding). To accomplish perceptually lossless coding, we need to make sure that the quantization noise remains below the masked threshold. The redundancy removal is accomplished by encoding the quantizer indices with lossless coding techniques (e.g., Huffman coding). To avoid confusion it should be clear that this is a lossless coding technique on quantizer indices; hence the overall operation is still a lossy coding

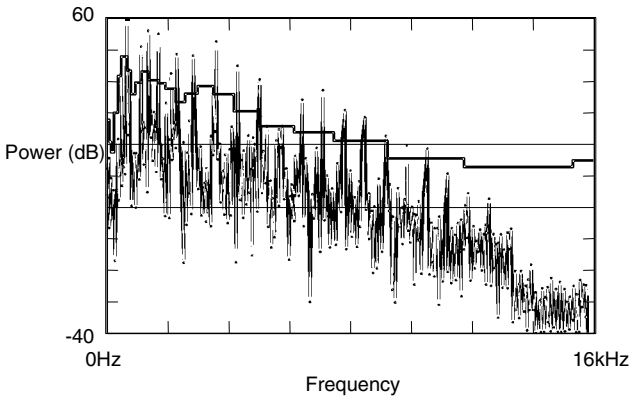


Figure 11: Example of signal frequency power spectrum and its corresponding masked threshold (solid stepped line).

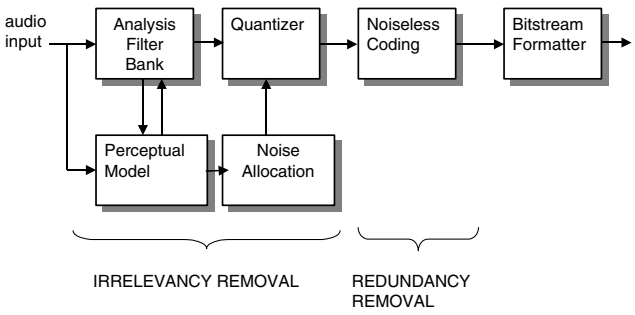


Figure 12: Block diagram of generic audio encoder.

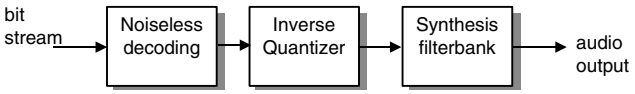


Figure 13: Block diagram of generic audio decoder.

operation. The resulting bit rate will be variable and will be signal-dependent. In many implementations an iterative procedure is used to find the optimum quantizer step sizes that result in coding noise below the masked threshold that will result in the lowest possible bit rate.

The decoder operation performs the operation in reverse, without the need for a perceptual model. A generic block diagram of an audio decoder is shown in Figure 13 . After the coefficients are reconstructed, the signal is transformed back to the time domain and ready for playback.

The block diagrams of Figures 12 and 13 are the principle for coding a single audio channel. For encoding multiple channels (N channels) once could in principle use N of these encoder/decoder pairs. However, in practice one would like to take advantage of the possible correlations that exist between the various channels. Also, for transparent coding, one would take into account that masking levels will differ for signals that are spatial in nature. Some distortions that are inaudible in each individual channel will become audible when listening to its multichannel version (e.g., in stereo).

Most state-of-the-art audio coders will produce a variable bit rate. In most communication scenarios a fixed bit rate is more desirable. This is accomplished by a buffering scheme that interacts with the coder quantization decisions. Designing buffering schemes that minimize buffer size (and its corresponding delay) and minimize impact on audio quality turns out to be a challenge, and various solutions exist, each with advantages and disadvantages.

Audio Coding Standards

Two types of standards exist. The first type is based on sanctioning by a standard organization (e.g., ISO and its MPEG standards). The second type is based on a proliferated proprietary *de facto* standard (Windows Media Player, RealPlayer, Dolby AC-3, etc). In both cases proper licensing is needed to use these standards in commercial applications. The MPEG coding standards are widely used for audio and video. The well-known MP3 standard is actually MPEG-1, Layer 3. Table 4 summarizes the MPEG standards.

The MPEG standards are defined in a different way than most ITU-T speech coding standards. They consist of a normative part defining the bit stream syntax and the decoder description (sometimes accompanied by reference C code). As a result anyone should be able to implement proper decoders. The encoder is described in the informative part and just describes the algorithmic concepts and operating modes. However, to build good encoders it is necessary to understand the algorithms, and the standard document will not provide details on how to build a good encoder. As a result standard compliant encoders can be quite different in performance.

Besides coders based on the traditional coding paradigm described above, other paradigms exist as well. These alternative paradigms play a role in very-low-rate coding. Two paradigms that have become part of the MPEG-4 standards are structured audio and parametric audio coding. Structured audio consists of a set of tools that can be used to generate music in a way similar to a music synthesizer. It also contains a structured way of describing musical scores and their subsequent conversion to sound using synthesizers. A structured audio bit stream describes how to build the synthesizers and provides the musical score and information on how to play this on the synthesizer. The resulting description can be very compact and low-bit-rate (only several kb/s). The resulting audio signal can be of very high quality, and because many modern music productions use synthesizers extensively, it can be very close to some originals. Parametric audio coding uses ideas similar to those used in speech compression by modeling some of the production mechanisms of the music sounds. Although it will work well for some sounds, it is difficult to make this technique work in a consistent way for a wide variety of input signals.

Table 4 Overview of Various MPEG Audio Standards

Standard	Year	Rates for transparency	Channels	Comments
MPEG-1 Audio, Layer I	1992	384 kb/s stereo	Mono, stereo	48-, 44.1-, 32-kHz sampling rates
MPEG-1, Layer II	1992	192—256 kb/s stereo	Mono, stereo	48-, 44.1-, 32-kHz sampling rates
MPEG-1, Layer III	1992	128—160 kb/s stereo	Mono, stereo	48-, 44.1-, 32-kHz sampling rates
MPEG-2, Layers I, II, III	1994	See MPEG-1 rates	Mono, stereo, and backward compatible 5.1 multichannel	MPEG-1 + enhancements, supports lower sampling rates
MPEG-2, AAC	1997	96—128	Mono, stereo, multichannel up to 48 channels	Not compatible with MPEG1, 2, supports 96-kHz sampling
MPEG-4 Version 1	1998	9—128	Mono, stereo, multichannel	Supports various coding tools
MPEG-4 Version 2	1999	9—128	Mono, stereo, multichannel	Support error robustness tools

APPLICATIONS

Two popular Internet applications of speech and audio compression are telephony and streaming. In both cases the IP data network is used to transport digitized (and compressed) audio signals. The basic protocol is the IP (Internet protocol), which is used to set up connections between machines and sessions between applications. For most data applications the next protocol layer would be TCP (transmission control protocol), which guarantees a reliable connection. Because this is accomplished by acknowledgment of receipt and possible retransmission it cannot guarantee continuous throughput and is therefore not suitable for telephony applications. Instead the UDP (user datagram protocol) is used (see Figure 14) . UDP is a simple protocol that does not support retransmission. As a result the data transmission is based on a best effort. For voice communications the use of UDP alone is usually not sufficient. For example, all packets could arrive but the order could have changed. Another protocol layer that is usually used on top of UDP is real-time transport protocol (RTP). This protocol is used to provide delivery services for real-time data such as time stamping, source identification, and sequence numbering. The sequence count is used to determine the playback order of packets and to identify if packets are missing. The time stamp is used to determine the amount of delay variation and could be used, for example, to adjust a jitter buffer. The source identifier is used to allow different calls to be combined into bigger packets. The RTP protocol is often used together with the real-time transport control protocol (RTCP). This defines a mechanism for hosts conducting a RTP session to exchange information for monitoring and control of packet counts, and number of packets lost. The RTCP packets use the same header as RTP packets, but its payload contains the control information. The RTCP packets are transmitted only several times per second to reduce overhead, but at least once every 5 s. RTCP can be used to monitor network performance, for example by reducing the load introduced by non-real-time traffic that is part of the same session (e.g., graphics) if too many packets are lost. It could also control the compression algorithm by increasing or decreasing its bit rate.

The use of the various protocols significantly increases the overhead (i.e., the extra information beyond the useable (payload) information). The currently common standard IPv4 has a 20- to 24-byte header, whereas the new proposed IPv6 header has a 40-byte header. UDP and RTP add another 20 bytes, resulting in a significant packet overhead. One way to minimize this overhead is by increasing the packet size. For audio signals this is not always desirable, because a large packet size corresponds to a large delay, or in case of packet losses, the removal of a large time segment of audio information.

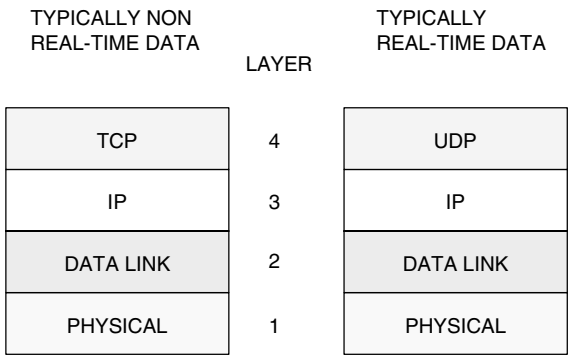


Figure 14: Protocol stacks for non-real-time and real-time sensitive data.

Another challenge in transmitting real-time sensitive data over a packet data network is the fact that the Internet is a best effort network. This means that one cannot guarantee the on-time arrival of the audio data, resulting in interruptions in the audio, or introducing unacceptable delays. In other words, *quality of service* (QoS) cannot be guaranteed for real-time communications. Many recent efforts have focused on improving this situation. For example, new protocols support differentiation of packets in terms of being real-time critical or not and improved routing schemes reduce the overhead needed for traffic flow management. To give real-time sensitive data a better guarantee on arrival it is necessary to prioritize these packets. It is important to keep in mind that prioritization can only work if other packets can get a lower priority and if routers can handle these priority requests. The resource reservations protocol (RSVP) tries to address this problem by negotiating priorities with routers. Another approach is based on the use of the differentiated services protocol (DiffServ), which sets priorities in the header bits and assumes multiple priority queues in the router. DiffServ is considered to be a more manageable protocol when used on a large scale, because it does not require routers to keep track of all requests and streams. Both protocols assume that the desired QoS result is achieved, but they will not guarantee it.

Internet Telephony

The use of the Internet for voice communications is now commonly referred to as *Internet telephony* or *voice over IP* (VoIP). Vocal Tech introduced the first consumer implementation in 1985 and it was widely hailed as a way to make free phone calls both domestically and internationally. As a result the concept of Internet telephony has become a reality, especially when used over managed networks such as found in enterprise networks. It not only will replace traditional circuit-switched phone systems in terms of functionality but also will open up many new opportunities. Examples are multimedia conferencing, integration of productivity tools on a PC with phone functionality, and Internet call centers. Economically, it makes sense to integrate voice and data services, not only from the end-user's perspective but also from the enterprise or service provider perspective.

To transport voice data over the IP network the signals are digitized and compressed with any of the coders described in the previous sections. The most commonly used ones are 64 kb/s ITU-T G.711 or lower rate coders such as the 8 kb/s ITU-T G.729 and G.729A. The media streams are transported using the IP/UDP/RTP protocol layering described before. To provide a complete phone service it is necessary to support numbering schemes, billing and call setup protocols, and servers that can acts

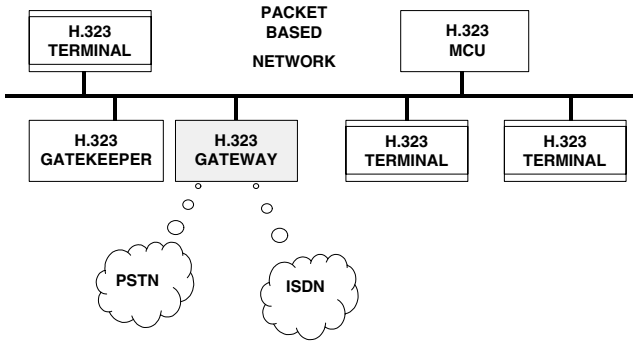


Figure 15: An example network topology using the H.323 protocol.

as intermediaries between end points. Two protocols that are widely used are H.323 and SIP. Both of these protocols build on existing protocols. The H.323 protocol suite was defined by ITU-T to support multimedia communications over IP independent of network topology. It contains four endpoints: (1) the terminal, which serves as the interface to the user, (2) the gatekeeper, which supports services such as billing and authentication, (3) the gateway, which supports connection to other networks, and (4) a multipoint control unit (MCU), which supports teleconferencing. Figure 15 shows an H.323 network topology. A simple end-to end call can be accomplished with two H.323 terminals, without the need for the functional modules.

Although we can build a telephony network independently of the traditional PSTN (public switched telephony network), in most cases we need to be able to interface with the legacy networks, for example when making a call from an IP phone to a traditional PSTN phone. The interface function between the Internet-based telephony and PSTN telephony is served by a so-called gateway. This gateway will facilitate the protocol translation and will also convert the audio data into a format suitable for the PSTN network.

The session initiation protocol (SIP) was proposed by IETF. It is modeled after the simple mail transfer protocol (SMPT). It is independent of the underlying packet protocol, and it leverages the Internet and Web structure. Although most existing IP phones use H.323, SIP is gaining momentum due to its simplicity. It is expected that both protocols will coexist for a long time and that most equipment will support both protocols.

Despite the use of RTP, there are still no guarantees that all packets will arrive on time, or will arrive at all. Buffering the incoming packets can compensate for late arrival, at the expense of an increase in delay. Hence it is common to make this buffer, the so-called jitter buffer, adaptive and to have it increase in size if the throughput becomes less reliable. Proper error mitigation is also important, and some coders have this built in, whereas others such as G.711 need external concealment. Due to buffering and other transport delay mechanisms, quite often there is a need for echo cancellation, especially if a speakerphone is used (or the speakers on a PC).

Although it is hard to believe that Internet telephony over the public Internet will replace PSTN services soon, it is a very likely to become a dominant application in corporate networks. Besides the economical advantage of only needing one network structure, these networks are also more carefully managed, thereby making it possible to guarantee a minimum quality of service.

Audio Streaming

There are two popular ways of distributing music over the Internet. *Downloading* is a copying operation from some server containing the original or compressed version of the material. The downloading will typically use the IP/TCP protocol, and one has to wait till all the material has been downloaded before being able to play the music. Once downloaded, the same material can be played repeatedly. The other common form of music distribution is *streaming*, which is very much like radio broadcasting. The material is transmitted to the end user, but the signal is played out almost instantaneously without local storage. Similarly to the VoIP scenario, it is necessary to buffer the packets to compensate for late arrival. Because this is a broadcast or one-way communication scenario, the amount of buffering can be large (a few to 30 s). Although in principle the buffers could be made even larger, this would create large delays before a player produced an audible signal. From a user interface perspective this is undesirable. Hence, sophisticated buffer control has been developed to make sure that continuous throughput is maintained, while maintaining a relatively small buffer. Some streaming services accomplish this by reducing the coder rate temporarily if the average available connection rate cannot support the initial streaming rate.

The compression techniques used are typically based on audio coding algorithms, because most material that is being streamed will be music. For lower rates it is possible that speech-coding algorithms are used instead. A commonly used format is the MP3 (MPEG-1 Layer III) format running at bit rates varying from 64 to 128 kb/s. For streaming applications, it is possible to use proprietary coders, as long as the decoders are available as downloads. However, most content providers will only support one or two formats, and as a result a couple of proprietary standards have become de facto standards. Examples are Apple's Quicktime, Microsoft's Windows Media Player, and RealNetworks' RealPlayer. All of these proprietary coders have a reasonable quality vs. bit rate performance, while trading off other parameters such as delay, complexity, and audio bandwidth. As should be clear from the previous sections, the best quality is obtained at higher bit rates (e.g., 96 to128 kb/s), whereas at lower rates (e.g., 24 to 64 kb/s) tradeoffs will be made by reducing the audio bandwidth, or even switching to mono. It should also be noted that even at the same bit rate and using the same format, differences in quality could exist due to the quality of the source material and the encoder used. For most streaming applications it is important that a variety of rates can be accommodated to support the various connection speeds. It is also important that the decoder have a relative low complexity to allow it to run on the host processor. With the advances in computing speed, this has become less of an issue. However, if these formats are used for downloading in portable players, complexity becomes an issue because it is connected to battery life and cost.

CONCLUSION

This chapter has explained the basics of speech and audio coding and its applications to Internet telephony and Internet streaming. It should be clear that especially for the lower bit rates, there is no single coder that is good for all applications, and that careful tailoring toward the application is important. Speech and audio coding is a mature field, with many available solutions. Based on our current knowledge and the various constraints that exist, it is expected that future developments in this field will focus less on compression efficiency and more on application-specific issues, such as scalability, error robustness, delay, and complexity.

GLOSSARY

Audio compression Reducing the amount of digital information to describe audio and speech signals, while maintaining audio quality as much as possible.

Digitization Creating a digital version of an analog signal by sampling and quantization.

Error mitigation Techniques used to reduce the perceivable impact of transmission errors of a compressed audio signal.

ITU International Telecommunications Union, which sets standards for global communications.

Linear prediction A technique commonly used to remove redundancies in a speech signal.

Lossless compression A compression technique that uses statistical properties of a digital representation to create a compressed version. After decompression the output signal will be equivalent to the input signal.

Lossy compression A compression technique that creates a compressed version of the input signal by taking into account that the decoded audio signal may sound different.

MPEG Motion Picture Experts Group, which sets worldwide standards for media compression.

Perceptually lossless compression A particular subset of lossy compression techniques, in which the differences between input and out signals are not audible.

Quantization Converting an analog value into a discrete digital value.

Quality of service A term commonly used in the context of Internet communications and broadcasting to describe consistency in delivering a particular signal quality.

Sampling frequency The rate at which analog signal samples are digitized. The value of the sampling frequency will be twice the highest possible frequency that can be contained in the signal.

Streaming Delivering audio signals over a packet network for continued broadcasting.

Transcoding Converting from one compressed format to another by decoding and encoding. If the coders used are lossy coders, the transcoding process typically introduces additional degradation.

CROSS REFERENCES

See *Data Compression; Video Compression.*

FURTHER READING

Bosi, M., Brandenburg, K., Quackenbush, S., Fielder, L., Akagiri, K., Fuchs, H., Dietz, M., Herre, J., Davidson, G., & Oikawa, Y. (1997). ISO/IEC MPEG-2 advanced audio coding. *Journal of the Audio Engineering Society, 45*(10), 789–814.

Faller, C., Juang, B.-H., Kroon, P., Lou, H.-L., Ramprashad, S. A., & Sundberg, C.-E. W., (2002). Technical advances in digital audio radio broadcasting. *Proceedings of the IEEE, 90*(8), 1303–1335.

ITU-T. Retrieved May 23, 2003, from http://www.itu.int

Kleijn, W. B., & Paliwal, K. K. (Eds.) (1995). Speech coding and synthesis. Amsterdam, The Netherlands: Elsevier.

Madisetti, V. K., & Williams, D. B. (Eds.) (1998). The digital signal processing handbook. Boca Raton, FL: CRC Press.

MPEG Audio Page. Retrieved May 23, 2003, from http://www.tnt-uni-hannover.de/project/mpeg/audio

Painter, T., & Spanias, A. (2000, April). Perceptual coding of digital audio. *Proceedings of the IEEE, 88,* 451–513.

Pohlman, K. C. (2000). *Principles of digital audio* (4th ed.). New York: McGraw–Hill.

Quatieri, T. F. (2001). *Principles of discrete-time speech processing*. New York: Prentice Hall.

Standards and Protocols in Data Communications

David E. Cook, *University of Derby, United Kingdom*

INTRODUCTION

This chapter focuses on the importance of open standards in the development of today's interoperable and global Internet. Many definitions are possible but this from the ISO meets our needs:

> Standards are documented agreements containing technical specifications or other precise criteria to be used consistently as rules, guidelines, or definitions of characteristics, to ensure that materials, products, processes and services are fit for their purpose. (ISO, 2002)

Global standards protect users from incompatibility problems between competing suppliers and government-sponsored systems.

While providing an introduction to the most important standards bodies and their historical background in the world of data communications it also discusses the new and emerging multimedia-dominated applications. Several standard protocols for file interchange, in particular those associated with the Internet and electronic commerce (e-commerce), are covered in detail. A variety of applications have been spawned specifically for the Internet as it has matured; these vary from the functionality required for good management to the World Wide Web (WWW). The chapter covers the new and emerging standards associated with e- and m-commerce (mobile commerce). Currently the demand for wireless applications is creating a considerable number of variations on existing protocols and several emerging new ones. These later applications are dependent upon a variety of multimedia approaches to file interchange that have become increasingly popular as Web services. Originally the standards were driven by forces requiring data exchange as a consequence of their businesses; today businesses solely dependent on the presence of exchangeable data have emerged, and it is these that now demand new standards for interoperability. These businesses are those that we now collectively know as e-commerce (Aaron, 1997). They in turn are attempting to accommodate and assimilate into their structures new technologies of a wireless nature under the banner of m-commerce. Therefore, the future of e-commerce on the Internet requires the development of more efficient standards in five areas: security, multimedia, document control standards, banking (e-money), and associated enabling technologies.

Security is still the single most formidable block to the growth of e-commerce. Consumer fears in this area are well founded. Most of these fears are due to Internet fraud and not the inability of technology to provide standards for secure transactions. The use of the secure socket layer (SSL) standard is commonplace and its details are covered elsewhere in this book; however, we briefly consider the major security standards.

Multimedia applications are the key to many online sales and, given increasing availability of bandwidth for mobile devices, this area will grow rapidly. This chapter examines the modifications necessary to the older standards as well as the current developments, particularly MHEG, driven by the Working Group 12 (WG12).

Inevitably the increasing use of mobile applications is forcing the introduction of specific protocols and standards to cope with the rapid growth in this area.

Document control standards are now increasingly needed to cope with the potential for microtransactions across the Web. One of the related problems is how to recognize content providers contributions to multimedia content. This area is being tackled by the Joint Photographic Expert Group (JPEG) and Motion Picture

Expert Group (MPEG) groups. Independently subsets of the standard generalized markup language (SGML) are being developed for electronic document and data-mining purposes.

Electronic banking is not just about the ability to access an account but the need to automatically and securely record transactions, which are increasingly becoming smaller in size and more numerous. The problem here is one of cost efficiency per transaction and the protocols and standards used cannot themselves be allowed to make the transaction prohibitive.

Enabling technologies for the future of the Internet are those concerned with the explosive growth in mobile and wireless applications and subsequently the use of personal networks. Without adequate standards growth will continue to be sporadic.

It is with these emerging technologies that we look to the future and bring the chapter to a close.

The evolution of standards has historically been ad hoc and in the field of communications is known to date back to the Greeks; however, we are not concerned with these early signaling systems but with the advent of electronic communications in the past century. Initially developed as internal mechanisms for developing the owners marketplace the standards have grown in influence from the recognition of the global need to cooperate as corporate bodies outgrew their national boundaries and were forced to allow standards bodies to grow and act independently.

In order to understand the need for standards and the consequent proliferation of standards bodies it is instructive to look at the nature of telecommunications traffic in general. There are fundamentally two contexts, voice and data, with multimedia being a mix of the two. Each of these is carried on a network that may be of two types, circuit switched, e.g., telephone, or packet switched, e.g., the Internet. The consumer end may be fixed, mobile or wireless. This is further complicated by the capabilities of the technologies available at a given time. This was understood by the ITU, who classified into three "generations" the technological advances (IMT-2000, 2000). First generation (1G) is the traditional analog systems, second generation (2G) systems are digital and cellular, and third generation (3G) systems are digital and cellular with higher data transfer rates. At each stage and for each type suitable recognized standards must be available to the service providers so that they can be assured of the structure of their marketplace.

STANDARDS BODIES

No standard will survive for long if there is no organization to support and promote its use. Some of these are supported by government funding but a surprisingly important sector is not. These are voluntary and unpaid collections, not only of companies but also scientists, researchers, and other interested parties. There are several of primary importance to the data communications world and particularly the future of the Internet; this section discusses these key organizations. Perhaps the most important are the ITU and the International Organization for Standardization (ISO).

International Organization for Standardization (ISO)

Established in 1947, the ISO is a collection of national standards bodies from around the world. Each country is allowed one representative standards body as a member and shares in the running costs of the organization. The agreements reached within the ISO are published as agreed international standards (the initials ISO are not an acronym but come from the Greek word "isos," or equal). All technical fields are covered (there are approximately 2850 committees) by ISO except electrical and electronic engineering, which is the responsibility of the International Electrotechnical Commission (IEC). Joint committees formed with the IEC produce standards in the information technology fields.

The Internet

The Internet is extraordinary in the sense that it began (and mainly continues) as a collection of voluntary and, in the main, open bodies collectively determining the current direction and future of the most important networked system in the world today. The Internet Society (ISOC), which was established in 1992 as the need for some control was appreciated, consists of professional Internet experts that concern themselves with policies and oversee a variety of boards and task forces dealing with Internet issues (ISOC, 2002).

Several key associated groupings exist such as the Internet Engineering Task Force (IETF), the Internet Engineering Steering Group (IESG), the Internet Architecture Board (IAB), and the Internet Assigned Numbers Authority (IANA). The IETF is the protocol engineering and development arm of the Internet Society. Although it had existed for some time, the group was formally established by the IAB in 1986. It is this group that proposes protocols for consideration as standards, although it is possible for any organization to do so. This is due to the use of the requests for comments (RFC) process. The technical management of IETF activities is conducted by the IESG, which also has responsibility for the Internet standards process. It guides the standards process according to the procedures that have been published by the ISOC. The IAB, originally known as the Internet Activities Board, first came into being in 1983. It was the guiding force behind the organized approach to the Internet. Under the IAB both the IETF and the IRTF (Internet Research Task Force) were formed in 1986. It was reconstituted in 1992 as the Internet Architecture Board and now serves as the technology advisory group to the Internet Society and as such is responsible for defining the architecture of the Internet. IANA is in charge of all IP (Internet protocol) addresses and any other parameters defined for the Internet. A major part of its task is to ensure the uniqueness of such parameters.

In order to ensure that network research and development was conducted in an open manner the RFC process was formulated in 1969 as a result of the ARPANET development commissioned by the DoD (Department of Defense, USA) for research into networking. Each RFC is assigned a unique identifier no matter what its content. Some were just for fun, e.g., RFC 527 and RFC 968;

however, most are of a serious nature so that the system exists to this day. Of special interest here are those RFCs concerned with standards and codes of practice. The standards process begins with a specification, which undergoes several review phases within the Internet community. These reviews are usually conducted via mailing lists. Each version of the specification is published as an RFC. When the specification becomes a standard it keeps the same RFC but additionally has a label of the form STDxxx. A similar process applies to codes of practice where the label is BCPxxx (best current practice). Each standard must pass through a set of three maturity levels known as the "standards track." They begin life as a "proposed standard," at which level it is assumed they have no known problems or omissions. The second stage requires the specification to have two implementations from two different code bases. Once sufficient successful operational experience has been gained, the level is raised to that of "Internet standard."

Once approved, the RFC editor is informed and it is the editor's responsibility to ensure publication and currency.

The World Wide Web Consortium (W3C)

The World Wide Web is without doubt the most popular application ever to be placed upon the Internet. It consists of multitudes of files scattered across the globe on millions of computers. The Web began life at the European Particle Physics laboratory in 1989 when Tim Berners-Lee wished to exchange data with fellow scientists (Berners-Lee et al., 1994). In 1994 he made agreements between MIT and CERN to develop the Web with support from the European Commission and DARPA. The W3C collects ideas from around the world to form a view of future directions for the WWW. It then designs appropriate Web technologies and contributes to standardization through specifications (recommendations). All of this is offered freely as part of the WC3 stated philosophy. This raises ethical questions when companies attempt to commercialize areas of the Web or its infrastructure. W3C has over 30 working groups working on what it terms as "activities." Their results form the basis for recommendations, some of which are PNG (1996), CSS (1996), XML (1998), DOM (1998), SVG (2001), XHTML (2001), and XML-Signature (2002). For a complete list the reader is referred to the W3C Web site (W3C, 2002).

ATM Forum

Although standards for ATM have been in place since 1984 under the auspices of the ITU-T it was not until 1991 that the ATM forum was established to promote this broadband technology as the packet-switching technology for high-speed data exchange. (ATM is particularly suited to multimedia applications because it allows different data rate streams to operate concurrently.) Like so many of its counterparts the Forum is an international voluntary organization that makes "recommendations."

International Telecommunication Union (ITU)

The ITU is the world's oldest international organization, being first established on 17th May 1865 to manage the early international telegraph networks in Europe. Initially there were 20 European founding states in the International Telegraph Union (ITU); its current name came into effect on 1 January 1934. In 1947 under an agreement with the United Nations, it became a UN specialized agency with headquarters in Geneva. In 1992 the organization was divided into three sectors: Telecommunication Standardization (ITU-T), Radiocommunication (ITU-R), and Telecommunication Development (ITU-D).

The work relevant to the Internet is carried out by study groups that develop new and revised ITU-T recommendations. The ITU-T currently produces some 210 of these recommendations each year.

Commitée Consultatif International Téléphonique et Télégraphique (CCITT)

The CCITT was a separate organization that set international communications standards; it is now part of the ITU. Amongst its more well-known standards are those for fax transmissions, e.g., Group 3 and Group 4, and the V series for communication at lower modem speeds. With respect to the Internet, the CCITT provided X.25, a packet-switching protocol for WANs; X.400, the protocol for e-mail; X.500, an extension to X.400 that defines addressing formats so that e-mail systems can be linked up; and the I xxx series for user interfaces.

INTERNETWORKING MODELS AND LAYERING PROTOCOLS

Standards are necessary to provide interoperability between two network-connected hosts. For this to occur, both hosts must agree on their approach to issues such as call establishment, error correction, and data transfer. This agreement then forms a set of rules that defines a protocol that each peer layer must follow. These rules are formalized as standards, and in order to provide a common framework for these standards a reference model is defined. This section examines the common reference models.

OSI (7-Layer) and TCP/IP Models

The most widely known reference model is that of the Open Systems Interconnections (OSI), referred to commonly as the 7-layer model. It is most often compared directly to the transmission control protocol/Internet protocol (TCP/IP) model and we shall follow the same pattern in this section. For any model the concept of a service provider and receiver is used. Each standard divides the reference model into a number of layers, each of which acts as a service receiver or provider, depending upon the logical state it is in. The OSI reference model was produced by an ISO working group and is occasionally referred to as the ISO model. It is the generic model for all networked systems although other protocol standards use a different number of layers.

In the OSI 7-layer model the following layers are present: physical layer, data link layer, network layer, transport layer, session layer, presentation layer, and the application layer. It is common practice to use the numbers associated with these layers rather than their names;

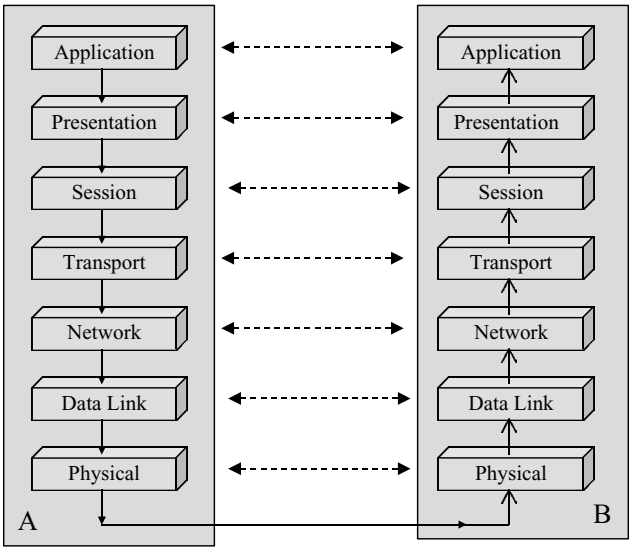


Figure 1: The OSI 7-layer model as applied to two networked hosts, A and B.

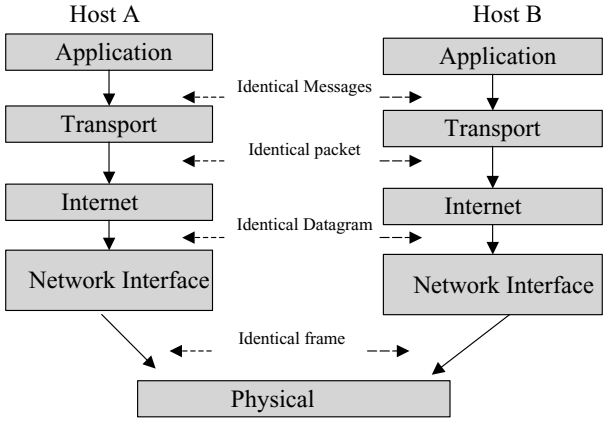


Figure 2: The reference model for TCP/IP. Note the lack of the session and presentation layers.

therefore the physical layer is known as layer 1 and so on. These interact with their corresponding layers across the network on a peer-to-peer basis so that each believes it has direct contact with its own peer layer. This is shown in Figure 1 for a single connection. The fact that each layer conceptually responds to its neighboring layer at the same level allows network connections to operate without the upper layers.

The physical layer (layer 1) defines the physical and electrical characteristics of a network interconnection. These are implemented in the wiring, the connectors, and the network interface cards of the hosts. The data link layer (layer 2) determines how the physical medium is accessed. HDLC, SLIP, and PPP protocols operate in part at this level. For the Ethernet MAC addresses are located at level 2 and this allows a LAN to become aware of its constituent members by their unique hardware addresses. The network layer (layer 3) allows systems using the same protocol to establish, maintain, and terminate connections. The IP protocol operates at the network layer, as do local hardware components such as routers. The transport layer (layer 4) ensures data reliability and integrity (TCP is usually considered to operate at this layer); retransmission requests and packet duplication problems are also resolved. The session layer (layer 5) has responsibility for ensuring sessions are completed without interference when necessary. This is important to e-commerce where the transaction cannot be disturbed by load sharing during its processing. The presentation layer (layer 6) deals with decryption, protocol conversion, and graphics expansion prior to presenting the information to the application layer. Layer 7, the application layer, is the part the user sees and uses. At this layer are applications such as FTP and other packages that require network access.

The TCP/IP structure is similar and due to its earlier commercial support (and support from the DoD in the USA) is the de facto applied standard. (It should be noted that the DoD did eventually formally move to the OSI model.) The structure is shown in Figure 2. See Figure 3 for the TCP/IP version of a network connection via a router (the OSI version is similar).

Each layer provides a recognizable service to the preceding and succeeding layers, which implies structural changes to the message stream if it is to be recognizable by peer layers. At any given layer we require transparency; e.g., each layer believes it alone is involved in the communications process with a peer layer elsewhere.

For the data link layer this requires it to know the physical address of the next destination or source, for the network layer it must know the network address, and for the user a pseudonym for the file is sufficient. This is achieved by a process known as encapsulation. See Figure 4. The original data packet from the TCP has been encapsulated within a datagram by IP, which in turn is enclosed within a frame. Depending upon the final protocol used for transmission at the data link and physical layers the structure may be further modified. The reader is referred to the appropriate chapter for the details.

It is evident that having a common network infrastructure provided by third parties will lead to further protocols and standards that apply to the physical layer of a real system. These parties operate their lines at a variety of speeds that determine the pricing for user access. Since

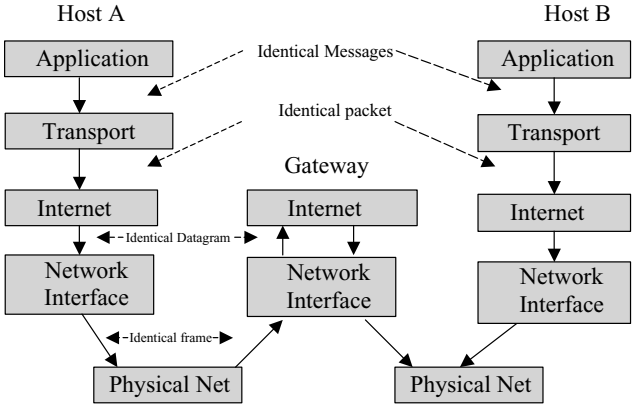


Figure 3: A simple TCP/IP Internet connection. Note the higher layers are not required.

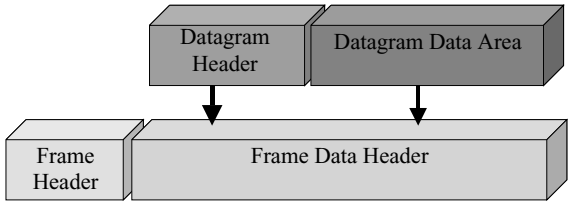


Figure 4: The original data are encapsulated within a datagram enclosed within a frame.

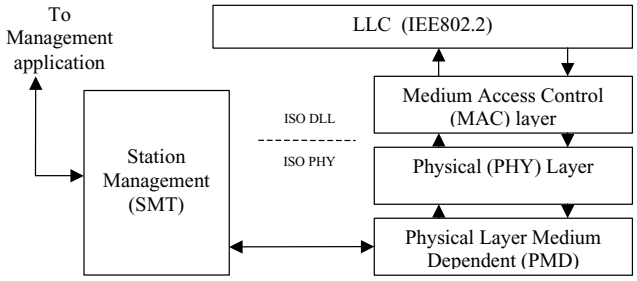


Figure 5: The FDDI reference model.

more than one user may access the same line a variety of multiplexing techniques are used. The following sections discuss some common protocols associated with delivery across nonproprietary networks.

Asynchronous Transfer Mode (ATM)

The ATM design was a response from the integrated services digital network (ISDN) design for a broadband mode to offer packet services to service providers across the digital subscriber networks. As such it is the delivery mechanism for B-ISDN services. It uses small packets or cells of 53 bytes for delivery: 48 bytes for data, 5 bytes for the header. This allows faster routers to be implemented but in order to reduce overhead a virtual connection approach is used.

The ATM reference model defined in CCITT I.321 introduces the concept of planes for different types of functionality, e.g., user, control, and management. Each plane is divided into layers where higher layers follow the OSI model but the lower layers, data link and physical, are replaced by the ATM adaptation layer (AAL), the ATM layer, and the physical layer.

The physical layer converts the ATM cell stream into bits for transfer across the chosen medium. Since ATM was aimed at broadband delivery it is usual to operate at 155.52 Mbps or the higher 622.08-Mbps rates delivered across fiber-optic cable.

The ATM layer provides multiplexing and demultiplexing of cells from different connections into a single stream, implementation of flow control, and management functions.

The ATM adaptation layer segments higher layer data into ATM size cells and vice versa. Several types of AAL have been defined to support different levels of service.

Fibre Distributed Data Interface (FDDI)

The fibre distributed data interface was designed as a very reliable high bandwidth fiber optic LAN. It is configured as a dual ring, thus ensuring continuity of service should one cable fail. Based on the original token ring protocol, IEEE 802.5, it was originally developed by the American Standards Institute (ANSI) working group X3T9.5 in 1982. It was not until 1988 that implementations first appeared. The FDDI reference model replaces the data link and physical link layers of the ISO model and connects to the IEEE 802.2 model logical link layer (LLC). Figure 5 shows the four layers peculiar to the FDDI, e.g.,the medium access control (MAC), the Physical, the Physical Layer Medium Dependent, and the Station Management layer.

The following is a brief description of the relevant layers:

The MAC layer controls network access and is defined in ISO 9314-2 "Information Processing Systems: Fiber Distributed Data Interface Part 2: Token Ring MAC."

The physical layer determines encoding and decoding and the clocking of the data stream to the line. Simply, it is responsible for the moving of the actual stream of 1's and 0's from host to host. It is defined by ISO 9314-1 "Information Processing Systems: Fiber Distributed Data Interface Part 1: Token Ring Physical Layer."

The physical layer medium dependent layer provides the actual connection to the ring and is defined in ISO 9314-3 "Information Processing Systems: Fiber Distributed Data Interface Part 3: Token Ring Physical layer Medium Dependent."

The station management layer functions are necessary for the control, supervision, and management of the connected stations. Its purpose at the network level is addressing and configuration.

Integrated Services Digital Network (ISDN)

The ISDN services were introduced in the 1980s to provide PSTN users with the opportunity to access services other than telephony. Initially converting analog service to digital it soon became apparent that a broadband service would be required (ISDN originally provided up to 2 Mbps, or 30 × 64 kbps channels). This was introduced as B-ISDN and initially used synchronous transfer mode due to technology limitations; in 1988 ATM was chosen as the carrier (see Asynchronous Transfer Mode (ATM)). It is the responsibility of the ITU CCITT group XV111.

TCP/IP PROTOCOL SUITES

The TCP/IP protocol is covered in detail elsewhere within this chapter. Also see Black (1995). Here we will briefly look at its origins and pay particular attention to its use and that of associated protocols on the Internet and in e-commerce. Although now universally known as two separate entities, TCP/IP began life in 1974 when Vint Cerf and Bob Kahn published "A Protocol for Packet Network Interconnection," which specified the design of a transmission control program (TCP) (Cerf & Kahn, 1974).

TCP/IP were split in 1978 and recognized as separate entities in 1982 when formally established as the ARPANET protocol. The IP has become the de facto standard for packet-switched transport across the Internet.

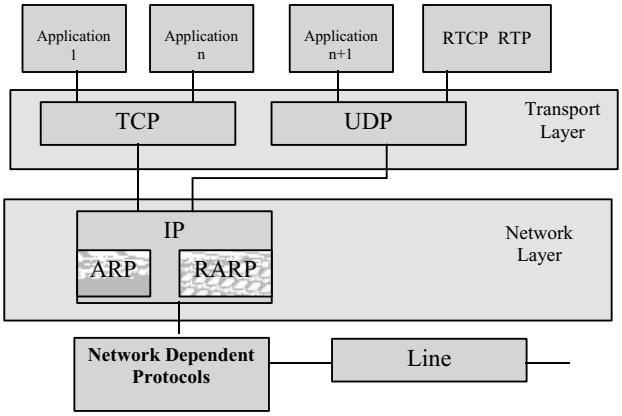


Figure 6: Relative levels of common Internet protocols.

Functionally, it is a connectionless protocol and relies on best-effort service. Although this provides a simple solution to a short message system over very reliable networks, it has problems when operating over public services.

Associated with the IP are several well-known service-providing protocols that operate at and above the network layer as defined by the ISO 7-layer model. They are the transmission control protocol (TCP), user datagram protocol (UDP), real-time transport protocol (RTP), and the real-time transport control protocol (RTCP). See Figure 6 for their relative positions in the reference model.

Transmission control protocol is offered as a connection-orientated service to the user at the transport layer. The user believes they have a permanent connection to the desired endpoint. The purpose of TCP is to provide application protocols with network-independent information exchange services.

User datagram protocol provides a connectionless service to application protocols. It is particularly useful for short message reporting.

The advent of the Web and multimedia content has provided a demand for real-time services across the Internet and hence via IP. The real-time transport protocol is able to transfer data types such as audio and video where real-time characteristics are required; however, it does not guarantee sequence delivery, error correction, or quality of service (QoS).

The real-time transport control protocol was designed for control of processes involved in conference calls. It can provide synchronization, framing, error detection, and source identification. It can also monitor the QoS. Both the RTP and RTCP protocols may be considered complementary and use separate ports; hence, they may be used together.

The Internet has grown as a collection of cooperative networks, each initially with its own addressing scheme. These became unified and where necessary correspondences between hardware addresses and Internet addresses are resolved and stored using the address resolution protocols (ARP) and the reverse address resolution protocols (RARP). Ultimately these addresses must be communicated to other networks so that data can be transferred. Thus devices known as routers are required. The routers calculate the most efficient path to a given address using the interior gateway protocol (IGP) within local groups. In order to check the availability of other routers outside the local group an exterior gateway protocol (EGP) is used.

Multicast and Unicast Control Protocols

It is often a requirement that more than one recipient receive a message or that more than one user are involved in a connection, e.g., conference calls. One such protocol is the Internet group management protocol (IGMP). IGMP is defined in RFC 1112. It is used to allow joining and leaving of multicast groups. Routers involved in multicasts periodically send membership queries to determine who is on the Net for that group. If a response is received, in the form of a report from a host detailing membership of groups, this is recorded for the forwarding of group messages.

IPv6 and the Future

It is only in the past ten years that the home user has become a major source of traffic on the Internet. This has led to an explosive growth in Internet address requests. The IETF has defined in RFCs 1883–7 a new version of the Internet protocol, IPv6, to replace the current version, IPv4. This has increased the address space to 128 bits from 32 bits. This new version addresses several other areas of concern. Amongst these are the introduction of improved security and data integrity, the ability of a host to acquire automatically an IP address without human assistance (this is particularly important to mobile applications), and better quality of service guarantees for multimedia applications.

INTERNET APPLICATIONS AND THEIR PROTOCOLS

The Internet is often seen as synonymous with the TCP/IP protocol suite and closely related protocols; however, many other protocols exist at various layers for specific purposes. This section looks at a few not covered elsewhere. It is interesting to note that DARPA in 1973 put the number of networks at over 7500. Currently according to figures obtained from the May 2002 Netcraft surveys (Netcraft, 2002) the number of active server sites now stands at 15.5 million. This level of use would not have occurred without the many applications that became available, the most famous of which is the WWW.

File Access (FTP, Telnet, NFS)

The Internet could not be the success it is without the early file access and transfer utilities and a supporting file structure standard known as the network file structure (NFS). Details of these application and support protocols may be found in RFCs 1122 and 1123. The file transfer protocol (FTP) is still considered the primary standard for file transfers on the Internet (as opposed to Web page delivery, which does not provide a local copy explicitly).

Remote access of files presupposes the ability to log onto the remote machine. This is catered for by Telnet, which is specified in RFC 854.Other log-in protocols such

as rlogin are available, and command shells such as rsh simplify access. More recently a secure shell known as SSH has become available. This ensures encryption is used to protect networks from hackers when logging on.

The World Wide Web (HTTP, HTML)

The original data transfer method for network use was FTP; however, with the advent of the WWW and the hypertext transfer protocol (HTTP), a different approach was possible. Now users could specify content for provision via a Web server as pages rather than files of material. HTTP therefore only needs to transfer the content of the page into the local Web browser for viewing. In this sense it is a one-way device, whereas FTP allows the file to be transferred, modified, and returned. Hypertext markup language (HTML) is the document-structuring language used by content creators on the WWW to create Web pages. It is a tag-structured language; although not strictly compliant as a subset of SGML, it is similar in concept.

The Next Generation (XML, SOAP, .NET)

The W3C is actively developing new features and services for the WWW, several of which are subsets of the SGML standard. Using HTML the content provider has restrictions on how they control their pages, which are set by HTML itself. Using the extensible markup language (XML) users can create their own tags. XML allows the definition, transmission, and interpretation of these data between applications. The proposed standard simple object access protocol (SOAP) was developed by Microsoft in conjunction with other software vendors as a way of adding document object models (DOMs) to standard HTTP streams. This effectively violates the purpose and intent of firewalls. Currently Microsoft is offering an OS known as .NET, which is specifically designed to operate as a distributed Web-based operating system. This will lead to the concept of paying for use of software as required since it will be downloaded from the relevant Web site on demand.

Emerging Multimedia Protocol Standards

Multimedia is now a common application format across all Internet-enabled devices; however, the needs of such data combinations of graphics, text, audio, and video are different from those historically associated with the Net. For an excellent introduction to the subject the reader is referred to Halsall (2001). Some of these needs are the following:

The audio and video must meet deadlines or defined time intervals or the user will experience indeterminate delays.

Variation in timings (jitter) must be bounded to avoid large buffers being needed to smooth the flow of data.

Multicast is a requirement for cooperative work and video conferencing.

Mechanisms are needed for synchronizing different data streams.

A current standard that addresses the needs of audio is known as MIDI.

The MHEG (Multimedia and Hypermedia Information Coding Expert Group) are one of several working groups of the ISO currently looking at standards for multimedia application use. The ISO/IEC JTC1/SC29 committee (Coding of Audio, Picture, Multimedia, and Hypermedia Information) carries overall responsibility for three working groups: WG1 Coding of Still Pictures (JBIG/JPEG), WG11 Coding of Moving Pictures and Associated Audio (MPEG), and WG12 Coding of Multimedia and Hypermedia Information (MHEG). Both the MPEG (Motion Picture Expert Group) and the JPEG (Joint Photographic Expert Group) are of particular importance to the multimedia world and the future of automated content costing for e-commerce use.

E-COMMERCE ENABLING PROTOCOLS

SSL/TLS and HTTP

Transactions over the Web for e-commerce require complete security if customer confidence is to be maintained. Without the assurance that others will not be able to access account details no potential e-commerce transaction will take place. Little can be done to stop certain types of fraud such as the nonexistence of goods for sale but the network transaction itself must be seen as secure. Several schemes exist but the secure socket layer (SSL) protocol, which lies in the session layer of the OSI model, provides authentication services for the client or the server. Full details of these services can be found in another chapter.

All e-commerce transactions require the transfer of documents of some kind. This implies that not only the content but also the structure should be exchanged. It is reasonable to require these structures, or document architectures, be the same at both ends of the network connection, hence the need for standards. The two current popular standards are the SGML and the open document architecture (ODA). A well-known application standard from typesetting is postscript. These document control systems have become increasingly important with the advent of the WWW and multimedia applications, and have coined the terms hypermedia and hypertext due to their abstract manner of access. The exchange format for multimedia is MHEG. Different task forces are working on the standards for hypermedia; amongst these are ISO/IEC JTC1 SC2/WG12 MHEG for coded representations of multimedia and hypermedia, the ANSI X2V1.8M for the Music Information Processing Committee, and the Standard Music Description Language (SMDL).

SGML (ISO 887) is a tag-based system, which means the standard specifies the form of the tags but user groups determine their position and meaning. The most famous application is HTML, which is understood across the WWW.

Emerging Mobile E-commerce Protocols

The ubiquitous use of mobile phones has provided the opportunity for Internet-based e-commerce applications to grow as an independent market area. Traditionally the main content for the telephone companies was voice; however, this has changed to data. A similar change is occurring with mobile traffic. Initially almost entirely voice,

the channels are now providing mixed voice and data channels. The Web has provided a mountain of content, which is a ready source for use in wireless communications. Many organizations now have Internet sites capable of supporting mobile applications, and wireless intranets are becoming more common. A typical use for these is in the reuse of historic buildings where new wiring is problematic or may not be allowed. A wireless intranet provides a cost-effective solution. This particular approach is currently being implemented in the University of Derby, England, at its Royal Devonshire Hospital campus in Buxton, which is governed by the specifications of the English Heritage Society. Efficient and common protocol standards are important if continued development is to be possible.

Currently the mobile phone services rely on the short message service (SMS) or WAP; however, the availability of wider bandwidth is encouraging the use of standards such as I-Mode from Japan, which is finding increasing favor in Europe. This is possible because national governments have made certain frequency allocations available for commercial use. Radio-frequency band usage varies from country to country due to historical reasons; however, they are sufficiently similar for standards to be set across national boundaries, for example, in the United States 2.4000–2.4835 GHz, in Europe 2.4000–2.4835 GHz (some variation), and in Japan 2.471–2.497 GHz.

The Institute of Electrical and Electronics Engineers (IEEE) instituted a working group known as IEEE 802.11 to create the current wireless LAN standards. The original standard had a signaling rate of 1–2 Mbps. Two variants rapidly followed, 802.11a and b. Surprisingly 802.11a operates at the higher frequency 5–6 GHz and has the greatest potential for supporting higher bandwidths (54 Mbps). 802.11b operates at 2.4 GHz and 11 Mbps. Each has their own supporters; however, 802.11b is rapidly becoming the accepted commercial standard.

802.11a products are currently expensive, and have high-power consumption. They are not compatible with 802.11b devices due to their different radio frequencies. 802.11b was named as Wi-Fi by the Wireless Ethernet Compatibility Alliance (WECA), an organization founded by commercial interests (3Com, Aironet (acquired by Cisco), Intersil, Lucent, Nokia, and Symbol) to promote Wi-Fi as a wireless standard. 802.11b products are generally backward compatible with 802.11 products. Several improvements have been made to security (802.11i), for use for multimedia (802.11e), and for higher bandwidths at the lower frequency range (802.11g). This latter works with the older wireless networking technology, 802.11b, allowing users to keep existing equipment and aimed at increasing the number of wireless connections in homes and businesses.

CONCLUSION

This chapter has tried to introduce the reader to the great variety of protocols and standards that have impacted the growth of the Internet. It is not possible to cover them all in the space provided; therefore, I have focused on those that have or will have the greatest impact on the development of the Internet. This chapter discussed the new and emerging standards associated with e- and m-commerce. This is an area where standards will continue to develop as the concepts of electronic shopping malls become more popular and e-banking becomes the norm. Wireless applications are in their Internet infancy and awaiting broader bandwidths. As this becomes available the scope for applications on a cost-per-view basis will increase. Of particular interest for the future are the attempts to commercialize WWW by offering software, which relies on the WWW's free infrastructure to be viable, on a pay-per-use basis. These in turn will need new protocols and standards to deal with transaction processing. These later applications are dependent upon a variety of multimedia approaches to file interchange, which is still an area of protocol research and development. Standards by their nature develop, mature, and die, to be replaced with versions more appropriate to their age.

GLOSSARY

CCITT (Commitée Consultatif International Téléphonique et Télégraphique) A separate organization that set international communications standards, but is now part of the ITU.

IEC (International Electrotechnical Commission) Produces standards in the information technology fields through joint committees with the ISO.

IETF The protocol engineering and development arm of the Internet Society.

ISO (International Organization for Standardization) A collection of national standards bodies from around the world that was established in1947, but is now often misnamed as the International Standards Organization.

ISOC (Internet Society) Established in 1992 and consists of professional Internet experts that concern themselves with policies and oversee a variety of boards and task forces dealing with Internet issues.

ITU (International Telecommunications Union) The world's oldest international organization, being first established on 17th May 1865 and is now a UN agency specializing in telecommunication standardization and development.

OSI An Open Systems Interconnections reference model that was produced by the International Organization for Standardization working group and is occasionally referred to as the ISO model.

W3C (World Wide Web Consortium) Collects ideas from around the world to form a view of future directions for the WWW, designs appropriate Web technologies, and contributes to standardization.

CROSS REFERENCES

See *Internet Security Standards; TCP/IP Suite.*

REFERENCES

Aaron, R., & Skillen, R. (Eds.). (1997). Electronic commerce [Special Issue]. *IEEE Communications, 40*(2).

Berners-Lee, T., Cailliau, R., Luotonen, A., Nielsen, H. F.,

& Secret, A. (1994). The World Wide Web. *Communications of the ACM, 37*(8), 76–82.

Black, U. D. (1995). *TCP/IP and related protocols*. New York: McGraw-Hill.

Cerf, V. C., & Kahn, R. E. (1974). A protocol for packet network connections. *IEEE Transactions for Communications, 22*(5), 637–648.

Halsall, F. (2001). *Multimedia communications*. Upper Saddle River, NJ: Pearson Education.

IMT-2000 (2000). ITU Study Groups. Retrieved from http://www.itu.int/ studygroups/areas-domain.html

International Organization for Standarization (ISO) (2002). *Introduction: What is ISO?* Retrieved May 30, 2002, from http://www.iso.ch/iso/en/aboutiso/introduction/whatisISO.html

Internet Society (ISOC) (2002). *All about the Internet*. Retrieved May 30, 2002, from http://www.isoc.org/internet/history

Netcraft (2002). Retrieved May 30, 2002, from http://news.netcraft.com/

World Wide Web Consortium (W3C) (2002). Retrieved May 30, 2002, from http://www.w3.org

Storage Area Networks (SANs)

Vladimir V. Riabov, *Rivier College*

INTRODUCTION

The volume and value of enterprise data have been growing faster than the speed at which traditional backup company utilities have been increasing. Enterprises have become more dependent on their online systems and cannot ensure the availability of these data by relying only on traditional, network-bottlenecked, server-attached storage systems. Solutions to this problem have been offered by SAN technology, which provides enterprises with serverless zero-time-windows backup. In this new backup approach, backed-up data are removed to a secondary remote storage device, and the enterprise server becomes off-loaded, permitting high-performance continuous access to both applications and data. A SAN fabric is usually based on fibre channel technology that allows up to 10-km long-distance connections. This feature has a significant advantage in a campus environment where reliable backup resources can be shared among several divisions. For example, fibre channel SANs for the Healthcare Enterprise allow 24 × 7 continuous operation, patient record backup, and medical image archiving (Farley, 2001).

The SAN becomes a key element of the enterprise environment where data availability, serviceability, and reliability are critical for a company's business. Many enterprise solutions (e.g., ATTO FibreBridge products, rack mount solutions, ATTO FibreCenter3400R/D, host bus adapters, the ATTO Diamond array, Compaq StorageWorks products, EMC Connectrix solutions, LSI Logic E4600 Storage System) are available today. They can be effectively used in multiple platform storage-infrastructure solutions for data-intensive applications such as e-commerce, online transaction processing, electronic vaulting, data warehousing, data mining, Internet/intranet browsing, multimedia audio/video editing, HDTV streaming, and enterprise database managing applications.

This chapter describes fundamentals of storage area networks (SANs), their architectural elements (interfaces, interconnects, and fabrics), technologies (fibre channel-arbitrated loop transport protocol, Brocade's configurations, InfiniBand switched-fabric architecture, crossroad systems with a storage router, virtual interface architecture, direct access file system, IP storage technologies, SANs over IP, fibre channel over IP, Internet fibre channel protocol, Internet SCSI, storage over IP, fabric shortest path first protocol, storage resource management, and adaptive network storage architecture), solutions, standards, associations, initiatives, forums, coalitions, vendors, and service providers.

SAN FUNDAMENTALS

What Is a SAN?

A SAN (*storage area network*) is a networked high-speed infrastructure (subnetwork) that establishes direct access by servers to an interconnected group of heterogeneous storage devices such as optical disks, RAID arrays, and tape backups, which are effective for storing large amounts of information and backing up data online in e-commerce, online transaction processing, electronic

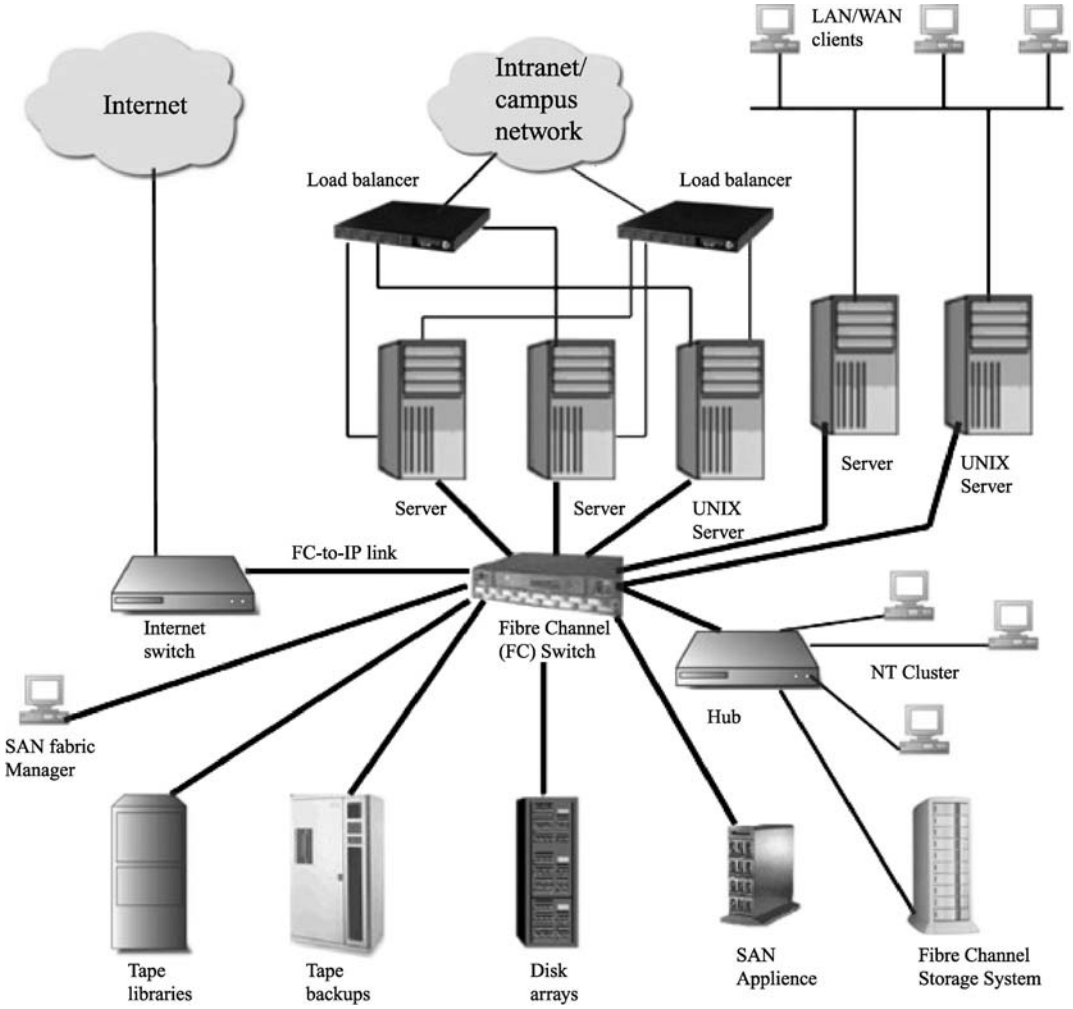


Figure 1: Storage area network as a networked high-speed enterprise infrastructure.

vaulting, data warehousing, data mining, multimedia Internet/intranet browsing, and enterprise database managing applications. SANs provide additional capabilities (fault tolerance, remote management, clustering, and topological flexibility) to mission-critical, data-intensive applications. A SAN is typically a part of an enterprise network of computing resources (Sachdev & Arunkundram, 2002). A simple model of the storage area network as a networked high-speed infrastructure is shown in Figure 1.

A SAN can be considered as an extended and shared storage bus within a data center, consisting of various storage devices and specific interfaces (e.g., fibre channel, ESCON, HIPPI, SCSI, or SSA) rather than the Ethernet (Peterson, 1998). In order to be connected to the enterprise network, the SAN utilizes technologies similar to those of LANs and WANs: switches, routers, gateways, and hubs (see Figure 1). Wide area network carrier technologies, such as asynchronous transfer mode (ATM) or synchronous optical networks, can be used for remote archival data storing and backup. As an important element of modern distributed networking architectures of storage-centric enterprise information processing, SAN technology represents a significant step toward a fully networked secure data storage infrastructure that is radically different from traditional server-attached storage (Clark, 1999). In addition, SANs provide improved options for network storages, such as a creation of remote or local, dedicated or shared data storage networks with access capability faster than *network attached storage* (NAS).

SANs are based on the storage-centric information-processing paradigm, which enables any-to-any connectivity of computers (servers) and storage devices over a high-speed enterprise network of interconnected fibre channel switches that form the SAN fabric. The incidence of unconnected clusters of information is eliminated or significantly reduced by SANs. According to this concept, a SAN resides behind the server and provides any users or devices on the enterprise network ("clients") with fast access to an array of data storage devices. It can be viewed as multihost connected-and-shared enterprise storage. Adding new storage devices and server elements resolves traditional network-bottlenecks and small-scale problems of interfaces, such as the small computer systems interface (SCSI) and network attached storage (NAS), and

can easily expand the scale of the SAN (Thornburgh & Schoenborn, 2001). Another advantage of SAN technology is that backups can be made over the SAN fibre channel subnet, and, in this case, backup traffic is totally removed from the enterprise network.

The SAN represents a new segment of the information services industry called storage solution providers (SSP). However, isolated SANs cannot realize SSPs' services, such as real-time data replication, failover, storage hosting, and remote vaulting.

Benefits of SANs

A SAN makes physical storage capacity a single, scalable resource and allows the flexible allocation of virtualized storage volumes (e.g., RAIDs, JBODs, and EMC, SUN, and DELL storage devices). The SAN can manage backup tasks that were a huge administrative and computer-resource burden under old storage architectures. The storage management cost savings can be higher than 80%. A cost-effective, scalable SAN enhances overall system performance. It can integrate legacy SCSI devices, which allows increasing their system-wide effective usable capacity by up to 30% (InfraStor, 2002).

SANs are an integral part of a large financial services enterprise, ISP, government organization, research laboratory, electronic publisher, digital video production group, TV-broadcasting station moving to digital services, or educational institution, or any organization with increasing data storage needs.

There are several key reasons for implementing a SAN (InfraStor, 2002). The first three concern business issues of return on the investment in data storage, as well as the protection of existing investments:

SANs are cost-effective (reduced cost of storage management, including backup and recovery; increased user productivity; cost-effective implementations of high availability disaster protection, using remote clusters and remote mirrored arrays);

SANs reduce business risk (faster disaster recovery; reduced revenue loss from down-time; reduced lost-opportunity costs);

Legacy investments are protected (SANs can be implemented without abandoning existing storage infrastructures such as devices using SCSI connections).

The next four address critical technical issues that face data-center managers at a time when the volume of data to be managed and made available in many organizations is increasing at a 60% annual rate (InfraStor, 2002):

SANs provide scalability (add servers and storage independently);

SANs allow flexibility (reconfigure storage and servers dynamically without interrupting their services; load sharing and redistribution);

SANs enhance overall system performance (more effective use of existing server compute cycles; real-time backup without impacting LAN/WAN; multiple server-to-storage paths; networked storage arrays that can outperform bus-attached storage; compatibility with parallelized database applications);

SANs are an integral part of any high-availability plan (facilitation of shared on-line spares and remote backup or mirroring; reduced down-time requirements; storage independent of the application and accessible through alternate data paths such as found in clustered systems).

The implementation of a SAN can realize significant overall cost savings in data-center operations and can increase user productivity. The opportunity to avoid escalating costs depends on decentralization of data and applications. A key element in the consolidation of data storage must include the implementation of a basic SAN infrastructure in order to provide (InfraStor, 2002) the following:

Bandwidth to service clients;

Maintenance of data availability without impacting LAN bandwidth;

Scalability for long term, rapid growth with protection of legacy investments;

Flexibility to provide optimum balance of server and storage capacity;

Manageability for ease of installation and maintainability;

Shared access to data resources for real-time backup and recovery.

Distributed environments require high-cost maintenance in terms of staff resources. The consolidation of distributed NT-based storages to a virtualized SAN-based resource can save 80% or more of the costs of management (InfraStor, 2002).

SAN Applications

SAN applications cover the following areas of data transfer (Peterson, 1998): (1) the externalization of data storage out of the server-SAN-attached-storage (SAS) and NAS-with-SAN-interconnects network architectures; (2) clustering, a redundant process that provides failover, high availability, performance, and scalability through the use of multiple servers as a data pipe and allows data storage resources to be shared; (3) data protection solutions for backup, remote clustering, file mirroring, and replicating and journaling file systems by creating data storage redundancy on a dynamic basis; (4) data vaulting, which is the process of transferring archived data to less expensive media; (5) data interchange from one storage system to another or between different environments; and (6) disaster recovery, which is similar to data interchange, moving copies of data offsite, and is built on remote vaulting (backup) processes or on remote array mirroring or clustering. Several new applications benefit from 2 Gb/s fiber channel SANs (Hammond-Doel, 2001): multimedia audio/video servers that provide the ability to stream higher resolution files, medical imaging, prepress that speeds up design and file preparation, and video editing of uncompressed HDTV data.

The first effective application of SANs has been serverless backup, which provides enterprises with full-time

information availability. All backup-related tasks have been relegated to the SAN. Large enterprises can store and manage huge amounts of information (several terabytes or more) in the SAN high-performance environment. Enterprise servers are connected to storage devices (e.g., RAID arrays) via a high-speed interconnection, such as fibre channel. The SAN any-to-any communication principle provides the ability to share storage resources and alternative paths from server to data storage device. A SAN is also able to share the resources among several consolidated servers.

A cluster of interconnected servers may be connected to common storage devices in the SAN environment and be accessible to all clients. Modern enterprises employ this clustering technology to resolve several challenging application problems (Barker & Massiglia, 2001, p. 244), i.e., providing customers, partners, and employees with *continuous* application service, even if the enterprise systems fail, and supporting application performance growth as demand grows, without service disruption to customers. Clusters provide load balancing, high availability, and fault tolerance and support application scaling. In some implementations, the clustered servers can be managed from a single console. Clustering methodology is effectively used in e-commerce, online transaction processing, and other Web applications, which handle a high volume of requests.

SAN methodology has its roots in two low-cost technologies: SCSI-based storage and the NAS-based concept. They both successfully implement storage-network links, but are limited to a low volume of data flows and rates. SCSI still remains the most popular "bus-attached" server–storage connection in SAN-attached storage (SAS) systems, especially at the stage of transition from SCSI bus devices to fibre–channel switches using the SCSI-fibre protocol converter in a new enterprise storage ("data center") environment. In the network attached storage (NAS) system, storage elements (i.e., a disk array) are attached directly to any type of network via a LAN interface (e.g., Ethernet) and provide file access services to computer systems. If the NAS elements are connected to SANs, they can be considered as members of the SAN-attached storage (SAS) system. The stored data may be accessed by a host computer system using file access protocols such as *NFS* or *CIFS*.

SANs provide high-bandwidth block storage access over long distance via extended fiber channel links. However, such links are generally restricted to connections between data centers. NAS access is less restricted by physical distance because communications are via TCP/IP (InfraStor, 2001). NAS controls simple access to files via a standard TCP/IP link. A SAN provides storage access to client devices, but does not impose any inherent restrictions on the operating system or file system that may be used. For this reason, SANs are well suited to high-bandwidth storage access by transaction-processing and DBMS applications that manage storage access by themselves. NAS, which has the inherent ability to provide shared file-level access to multiple OS environments, is well suited for such requirements as Web file services, CAD file access by combined WinNT/2000, UNIX, and LINUX devices, and wide-area streaming video distribution (InfraStor, 2001). A balanced combination of these approaches will dominate in the future.

SAN Architecture

The SANs architectures have been changed evolutionarily, adapting to new application demands and expanding capacities. The original fibre-channel-based SANs were simple loop configurations based on the fibre channel arbitrated loop (FC-AP) standard. Requirements of scalability and new functionality had transformed SANs into fabric-based switching systems. Numerous vendors offered different solutions of problems based on fabric switching. As a result, immature standards created various interoperability problems. Homogeneous high-cost SANs were developed. Ottem (Ottem, 2001) refers to this phase as the legacy proprietary fabric switch phase. The latest architectural approach is associated with a standards-based "Open" 2Gb fabric switch that provides all the benefits of fabric switching, but based on new industry standards (FC-SW-2) and interoperability architecture that runs at twice the speed of legacy fabric. The standards-based switches provide heterogeneous capability. The latest feature reduces prices of the SAN's components and management costs of running a SAN. Characteristics of three generations of SANs are summarized in Table 1.

The Open 2Gb fibre channel allows doubled SAN speeds, enables greater flexibility in configuring SANs for a wide range of applications, and is especially useful for managing 1.5-Gb high-definition video data. In the HDTV applications, a single fibre can carry a full high-definition video stream without having to cache, buffer, or compress the data. Other examples (Ottem, 2001) include storage service providers that must deliver block data from to users at the highest possible speeds and e-commerce companies that have to minimize transaction times. The 2-Gb fibre channel provides the high-speed backbone capability for fibre channel networks, which can be used to interconnect two SAN switches. This configuration increase overall data throughput across the SAN even if servers and disk subsystems continue to operate via 1-Gb channels.

A SAN system consists of software and hardware components that establish logical and physical paths between stored data and applications that request them (Sheldon, 2001). The data transforms, which are located on the paths from storage device to application, are the four main abstract components (Barker & Massiglia, 2001, p. 128): the *disks* (viewed through ESCON, FCP, HIPPI, SCSI, and SSA interfaces as abstract entities), *volumes* (logical/virtual disk-like storage entities that provide their clients with identified storage blocks of persistent/retrieved data), *file systems*, and application-independent *database management systems*. In a system with a storage area network, five different combinations (Barker & Massiglia, 2001) of these data transforms and corresponding transformation paths serve different applications and system architectures by various physical system elements. The disk abstraction is actually the physical disk drive. The abstract volume entity is realized as an external or embedded RAID controller, as an out-of-band or in-band SAN appliance, or as a volume manager

Table 1 Three Generations of SANs

	Fabric	Main Characteristics	Applications
First-Generation SANs	1Gb Loop	FC-AL protocol; 1Gb speed; enabled first SANs	SCSI replacement
Second-Generation SANs	1Gb Proprietary; Legacy Fabric	FC-SW protocol; 1Gb speed; proprietary switch-to-switch connections; expensive	LAN-free backup; HA clustering
Third-Generation SANs	2Gb Open Fabric	Open FC-SW-2 protocol; 2Gb speed; standards-based switch-to-switch connections; competition-driven price reductions price reductions	Serverless backup; heterogeneous storage consolidation; high-definition video; data; virtualization

serving a database or an application. Storage servers (such as NAS devices), database servers, and application servers may contain the abstract file systems. These devices and servers can be clustered to increase scaling and application availability. In that case, their volume file and management systems should be cluster-aware (Barker & Massiglia, 2001).

Any SAN-based client–server system consists of three architectural components: *interfaces*, *interconnects* or *network infrastructure components* (switches, hubs, routers bridges, gateways, multiplexors, extenders, and directors), and *fabrics*. The SAN interfaces are fibre channel, ESCON, HIPPI, SCSI, and SSA. The SAN interconnects link these storage interfaces together, making various network configurations. Routers and bridges perform protocol transformation in SANs. Switches increase the overall SAN bandwidth by connecting system elements for data transmission and allow advantages of the centralized storage repositories with the shared applications and central management. The most common SAN fabrics are switched fibre channel, switched SCSI, and switched SSA, all of which physically link the interconnects and determine the SAN's performance and scalability. Some fabrics embed operating systems that provide for SAN security, monitoring, and management. Hosts are connected to the fibre channel SAN through host bus adapters (HBAs), which consist of hardware and interface drivers. Fibre channel HBAs support negotiation with network-attached devices and switches and allow the host to minimize its CPU overhead.

SAN Operating System Software Components

The SAN software plays an important role in providing an environment for various business and management applications, called *system applications* (Barker & Massiglia, 2001, p. 13), such as clustering, data replication, and data copying. The *management applications* (zoning, device discovery, allocation, RAID subsystems, and others) manage the complex environment of distributed systems. These applications can significantly reduce the cost and improve the quality of enterprise information services.

SAN TECHNOLOGIES AND SOLUTIONS

The SAN infrastructures support multiple protocols, such as SCSI, SNMP, VI, ESCON/FICON, TCP/IP, and SSAIP, over a single physical connection. This unique capability provides the SAN system with a coupled functionality of an interface to storage devices and a server interconnect.

In the early 1990s, fibre channel was developed by the Fibre Channel Systems Initiative (FCSI) and adopted later by the ANSI X3T11 Committee as a high-speed interface for connecting storage devices to servers and other network configurations. These interconnect standards provide SANs with the vital properties of connectivity, bandwidth, interconnectivity, protocol efficiency, distance range, recoverability, failure tolerance, and cost options. The fibre channel standards specify electrical and optical transmission media, as well as conventions for signaling and transmission/functional protocols. Optical medium (with SC, LC, and MT-RJ connectors) supports reliable signaling over long distances. Fibre channel provides data rates in the range from 133 Mbit/s to 4 Gbit/s over low-cost copper cabling (shielded twisted-pair wire or coaxial cable with DB-9 and HSSDC connectors) or higher-cost multimode fiber-optic cable. Fibre channel fabrics have transceivers, called gigabit interface converters (GBICs), which convert optical to electrical signals to cable connectors. The fibre channel technology supports distances up to 10 km.

Fibre-Channel-Arbitrated Loop Transport Protocol (FC-AL)

The fibre channel methodology has means to implement three topologies: point-to-point links, arbitrated loops (shared bandwidth loop circuits), and bandwidth-switched fabrics that provide SANs with the ability to do

bandwidth multiplexing by supporting simultaneous data transmission between various pairs of devices. Any storage device on the loop can be accessed through a fibre channel switch (FCSW) or hub. The fibre channel switch can support entry-level (8–16 ports) to enterprise-level (64–128 ports) systems. Under the ANSI X3T11 standards regulation, up to 126 storage devices (nodes) can be linked in the fiber channel arbitrated loop (FC-AL) configuration, with the storage interface bandwidth about 100 Mbits/s for transferring large files. More than 70 companies, including industry-leading vendors of disk arrays and computer and networking systems, support the FC-AL voluntary standards. The FC-AL topology is used primarily to connect disk arrays and FC devices. Originally developed as the high-speed serial technology of choice for server–storage connectivity, the FC-AL methodology is extended to the FC-SL standard that supports isochronous and time-deterministic services, including methods of managing loop operational parameters and QoS definitions, as well as control. The FC-VI regulation establishes a fibre channel-virtual interface architecture (FC-VIA) mapping standard. See the chapter on fibre channels for more information about various implementations of the technology in various network configurations, including SANs.

Because of the high cost of the FC interconnect components and separation of storage and servers at the wide area network scale (resulting in slow capabilities of WAN–SANs with fibre channel), alternatives to FC technologies have been developed. The *ipStorage* technology (Barker & Massiglia, 2001, p. 187) employs TCP/IP as a storage interconnect. The Internet Engineering Task Force (IETF) has proposed the iSCSI (*Internet SCSI*) standards that address the issues of long distances (WAN-scale), reducing the interconnect cost, high security, and complex storage network topologies. The iSCSI is layered on top of the TCP/IP protocol hierarchy and can instantly access all modern transmission media and topologies. TCP/IP and related protocols have been implemented in the server-based systems that allow the most general storage networks to be constructed with the iSCSI methodology. The main challenge is a reduction of the iSCSI processor overhead of operating iSCSI packets below the Fibre Channel overhead level.

InfiniBand Solutions

InfiniBand is a new emerging interconnect technology, developed by the standards of the InfiniBand Trade Association (founded by Compaq, Dell, Hewlett-Packard, IBM, Intel, Microsoft, and Sun Microsystems) that offers the most general low-cost server system topologies (InfiniBand Trade Association, 2003). It is expected that InfiniBand interfaces will be embedded into all Intel-based servers (Barker & Massiglia, 2001, pp. 188–192; Intel InfiniBand Architecture, 2003) and will allow Windows and Linux servers to be available for resolving complex problems of data centers by adopting clusters and multihost enterprise RAID subsystems. The InfiniBand technology implements a switched-fabric architecture with the packet-switching communication protocol (PSCP) that relates to the virtual interface (VI) architecture methodology. SANs, parallel processing systems, and systems area networks can effectively use InfiniBand as a high-performance/low-latency interconnect.

Crossroads Systems With a Storage Router

InfiniBand technology has been successfully implemented by Crossroads Systems, Inc., which promotes storage solutions based on protocol-independent connectivity at Gigabit/s speeds and unparalleled manageability for various storage devices. The Crossroads' storage routers (e.g., Crossroads™ 10000) support peer operations between storage devices and multiprotocol servers on fibre channel storage networks.

Brocade's Configurations

Brocade Communication Systems, Inc., has developed an intelligent fabric services architecture that creates a scalable and secure environment for enterprise mission-critical storage applications such as data backup and business continuity. The Brocade SANs (SilkWorm™ family of fabric switches and software) provide enterprises with any-server-to-any-storage-device connectivity and consolidate storage resources and servers, as well as sharing backup resources (Beauchamp, Judd, & Kuo, 2002).

OTHER STORAGE NETWORKING TECHNOLOGIES

The following emerging technologies introduce new system architectural approaches in storage networking. SAN developers and users are trying to adapt them to a new enterprise environment that is characterized by host-level heterogeneous complexity, management flexibility, new TCP/IP network communication services, file-access-protocol developments, and the repartitioning of the functionality of the file management systems.

VI (Virtual Interface) Architecture

The virtual interface (VI) architecture is a midlayer protocol specification that regulates virtual intercommunication between applications running on different remote servers (i.e., in a cluster). This methodology significantly reduces the latency and the volume of the system I/O operations by using message and data buffer pools that are insensitive to heterogeneous operating environment or other applications. The reduction of the I/O-related interrupts increases the CPUs' time for processing various other system tasks. Developers of the VIA technology (Compaq, Intel, Microsoft, etc.) utilize this architecture as an efficient way of message communication between the SAN nodes at the application level, creating only a small overhead of intercommunication between the remote applications. This methodology has been successfully implemented in database managers and NAS devices (Barker & Massiglia, 2001, pp. 189–190). Several efforts (i.e., the Direct Access File System and Network File System initiatives) have been made to improve file system performance by utilizing VIA-type advanced transport protocol features. The Emulex Corporation promotes the GN9000/VI™ 1 Gb/s VI/IP PCI host bus adapter, which

is based on the virtual interface (version 1.0) architecture, supports standard TCP/IP-reliable data delivery, IP routing, and the direct access file system (DAFS) standard, and speeds data access over standard Gigabit Ethernet networks.

Direct Access File System (DAFS)

The direct access file system is a new file access/transfer protocol that is based on CIFS/NFS characteristics and VIA-type transport protocol features. DAFS/VIA technology supports direct file transferring between the storage system and clients. In the SAN environment, data can be directly transferred among a number of servers.

IP Storage Technologies

Another *block-mode data* mechanism has been used by the IETF IP Storage Working Group in developing standards for a new IP-based transport-through-network technology that encapsulates fibre channel and SCSI high-speed interfaces and provides direct access to data on disks, tapes, and optical storage devices. IP storage technology allows embedding low-cost SANs into IP-based enterprise infrastructures over existing Gigabit Ethernet networks.

SANs Over IP

To avoid the distance limitation of the fibre channel interconnects, enterprises build remote SANs that can be interconnected by means of the SAN-over-IP technology originally developed by the Computer Network Technology Corp. The distant SANs appear as local storage entities. This technology improves enterprise management and data access, disaster recovery, business continuity, disk mirroring, electronic tape vaulting, and wide area clustering. The Storage Networking Industry Association (SNIA) offers three technologies for integrating fibre channel SANs into the IP backbone. These methodologies include fibre channel over IP (FCIP), iFCP, and Internet SCSI (iSCSI). The FCIP, iFCP, and iSCSI transport protocol descriptions are presented in (Clark, 2002) and on the CNT Web site (CNT, 2002).

Fibre Channel Over IP (FCIP)

FCIP is the simplest point-to-point IP tunneling solution for intercommunicating remote SANs with fibre channel fabrics. The FCIP gateways establish TCP/IP connections over a WAN path to transport the fibre channel encapsulated frames. A typical discrepancy in data communication rates between an FCIP-attached WAN link and fibre channel fabric generates various flow control issues that can be resolved by TCP sliding-window algorithms. Several FC-FCIP management issues cannot be properly determined for the FCIP pipes because the TCP/IP transport component ends at the external nodes of the fibre channel network. These problems have been addressed and successfully resolved in the iFCP and iSCSI approaches.

Internet Fibre Channel Protocol (iFCP)

The gateway-to-gateway Internet fibre channel protocol (iFCP) supports a means of integrating fibre channel end devices into a single IP SAN. By using iFCP, the fibre channel fabric services can be provided to the remote FC devices over a TCP/IP network. The iFCP IP storage switches can directly connect fibre channel storage arrays, HBAs, hubs, switches, and routes. The iFCP is a protocol stack, which can be implemented in an IP storage controller interface or integrated into Gigabit Ethernet IP storage NIC (known as ANSI X3T10 and X3T11 Standards) (Clark, 2002, pp. 126–139). It supports any-to-any IP routing of storage data. A mismatch in data communication rates between an iFCP-attached WAN link and fibre channel fabric generates various flow control issues that can be resolved by TCP sliding-window algorithms. The IPSec, public or private keys, and zoning methods can provide security across the Internet. One of the important applications of the iFCP technology is the support of multiple TCP/IP connections for concurrent storage transactions.

Internet SCSI (iSCSI)

In contrast to the FCIP concept, the iSCSI methodology, which follows the SCSI client/server model, is based on the implementation of a light switch technology in IP storage networking (Clark, 2002, pp. 139–149) and excludes fibre channel elements. The iSCSI servers (*targets*) are present in disk arrays and client nodes (*initiators*) that occupy host platforms. The iSCSI protocol over the TCP/IP layer is used for block data transport between these entities over the IP network. Data can be directly written into application memory through a steering and data synchronization layer located below the iSCSI sublayer. IPSec, Keyberos, public key, and other methods can provide security across the Internet. SANs use the iSCSI adapters with TOEs to minimize processing overhead and realize high-performance features of the iSCSI technology. The enterprise solutions with IP SANs can also support the Gigabit and faster Ethernet on iSCSI-switch infrastructures.

Storage Over IP (SoIP)

Based on the SoIP remote storage technology, the Nishan Systems Corporation developed IP Storage switches of the IPS 4000 Series™and a suite of storage management software that allow configuration and monitoring of large-scale storage networks.

Fabric Shortest Path First (FSPF)

Fabric shortest path first (FSPF) is the OSPF-based standard routing protocol for fibre channel that determines the next shortest route for data traffic, updates the routing table, and detects the failed routes (Vacca, 2002, p. 152). The optical, link, or switch failures can be effectively handled by FSPF with minimal impact on the interconnected devices in the FC/SAN environment.

Adaptive Network Storage Architecture (ANSA)

The Procom Technology Corporation has developed the adaptive network storage architecture (ASNA) approach, which delivers both block level and file access to data.

Procom's NetFORCE3000™Series provide filer functionality (together with advanced features of security, high stability, backup, and recovery) to an enterprise storage that can result in high-performance information management systems. The ASNA technology has been successfully applied to database management, data warehousing, e-mail delivering, and 24 × 7 rich media applications.

Storage Resource Management (SRM)

The SRM technology provides applications for managing logical and physical storage-system resources (virtual devices, disk volumes, file resources, storage devices and elements, and appliances). SRM tools allow storage-system administrators to configure and monitor SANs and other storage resources. During the administrative monitoring, the transport or storage data remain unchanged. Vendors of the SRM tools, products, and services include SUN Microsystems (*Sun StorEdge*™), HighGround Systems, Inc. (*Storage Resource Manager*, *Media Mirror*), and Storage Computer Corp. (*Storage Administrator*) (Toigo, 2001).

STANDARDS

American National Standards Institute (ANSI)

The American National Standards Institute (ANSI) coordinates SAN voluntary standards (ANSI, 2003). The ANSI X3T10 and X3T11 working committees are associated with storage networking issues including SCSI I/O interface standards (X3T10) and fibre channel interface standards (X3T11). The first set of fibre channel standards (ANSI X.3230–1994) (Vacca, 2002, pp. 75–78) describes standards for a switch fabric (FC-SW2), the interconnect that supports high volumes of throughput and bandwidth for disk output and input, as well as a management information base (MIB) management standard that permits fibre channel devices (switches) to be managed by any vendor's software, which includes an implementation of the simple network management protocol (SNMP). SAN users can find a brief description of other X3T11 Fibre Channel standards in (Barker & Massiglia, 2001, pp. 384–386).

Several FC SAN equipment vendors, including Brocade Communications Systems, have refined SAN standards in the areas of management, discovery, data transport, and WAN connectivity. This allows the fibre channel SAN to become an integral part of the enterprise framework (Vacca, 2002, pp. 78–79).

Distributed Management Task Force (DMTF)

The Distributed Management Task Force (DMTF) has introduced management standards for computer systems and enterprise environment (DMTF, 2003). The SAN-related management standards (Barker & Massiglia, 2001, pp. 386–387) cover a set of the Web-Based Enterprise Management (WBEM) XML-based technologies. They support an object-oriented approach in developing a business's management environment, using the Common Information Model; architectures and frameworks for desktop, laptop, and server management (desktop management interface); standard data models for a network, its elements, policies, and rules (directory enable networks); and functional calls for transaction monitoring (known as the application response measurement [ARM] standard).

Storage Systems Standards Working Group (SSSWG)

As a division of the Institute of Electrical and Electronics Engineers (IEEE), the Storage Systems Standards Working Group (SSSWG) develops models and architectures of storage systems, including SANs (SSSWG, 2003). The SSSWG project authorization requests (Barker & Massiglia, 2001, pp. 387–388) include the Guide to Storage System Design; Media Management System (MMS) Architecture; Session Security, Authentication, Initialization Protocol (SSAIP) of the MMS; Media Management Protocol (MMP) for both client and administrative applications; Drive Management Protocol (DMP) of the MMS; Library Management Protocol (LMP) of the MMS; the Media Manager Interchange Protocol (MMIP) for information exchange between autonomous Media Managers; the Media Manager Control Interface Protocol (MMCIP); the C Language Procedural Interface for implementation of the MMS's components; MMS User Mount Commands for establishing "command line interfaces" (CLI); MMS standard administrative and operational commands for administering and operating an MMS; and "MOVER" specifications of a storage system data mover architecture and its interfaces.

Internet Engineering Task Force (IETF)

The Internet Engineering Task Force (IETF) organization defines a variety of the transmission control protocol/Internet protocol (TCP/IP) standards that are widely used in the enterprise environment with SANs (IEFT, 2003). The IETF standards related to storage networking (Barker & Massiglia, 2001, pp. 388) include the simple network management protocol (SNMP) for managing and monitoring devices and systems in a network; the Internet protocol over fibre channel (IPoFC); and policy for quality of services (QoS).

STORAGE NETWORKING ASSOCIATIONS, INITIATIVES, FORUMS, AND COALITIONS

The following organizations promote storage networking technologies and products, develop standards, undertake marketing activities in information technology industry, educate, train, and create the knowledge base for implementing SAN technology:

SNIA (Storage Networking Industry Association)

The SNIA, an international association of developers of storage and networking products, is focusing on the creation of a forum of IT companies, system integrators, and application vendors for delivering architectures,

education, and services in storage networking, as well as defining the specifications and infrastructures, and proposing standards for storage networking systems, including SANs, SAN attached storage (SAS), and network attached storage (NAS) (SNIA, 2003).

Fibre Channel Industry Association (FCIA)

The FCIA is an international organization of manufacturers, systems integrators, developers, systems vendors, industry professionals, and end users. In June 2002, this organization included more than 190 members and affiliates in the United States, Europe, and Japan. The FCIA is committed to delivering a broad base of fibre channel infrastructure to support a wide array of industry applications within the mass storage and IT-based arenas. FCIA working groups focus on specific aspects of the technology that target markets, which include data storage networking and SAN management. The overview of fibre channel's SAN and networking applications and examples of fibre channel solutions for high-performance networks of heterogeneous storage, server, and workstation resources can be found on the technology section of the FCIA Web site (FCIA, 2003).

Fibre Alliance (FA)

The Fibre Alliance is the networking industry consortium originally founded by a group of storage networking companies, including EMC Corporation, to develop and implement standards for managing heterogeneous fibre-channel-based SANs. In collaboration with the Internet Engineering Task Force (IETF), this group develops the definition of simple network management protocol management information bases (SNMP MIB) for storage network and device management (Fibre Alliance, 2003).

Jiro

Jiro is a Sun Microsystems technology that delivers intelligent management services for networked devices. Using the principles of Java and Jini platform-independent application development interfaces (Jini, 2003), Jiro technology provides the architecture for connecting and managing complex distributed environments such as storage area networks. The Jiro technology brings higher levels of interoperability, adaptability, and manageability to enterprise networks with storage resources (Jiro, 2002).

National Storage Industry Consortium (NSIC)

Since April 1991, the National Storage Industry Consortium has consolidated the efforts of over 50 corporations, universities, and national labs in the field of digital information storage. The corporate members are major information storage manufacturers and companies from the storage industry infrastructure, including SANs. As a nonprofit organization, NSIC supports precompetitive joint research projects, involving collaboration among users and integrators of storage systems, storage system and device manufacturers, storage component and media manufacturers, suppliers, universities and national laboratories (National Storage Industry Consortium, 2003).

SANS' MARKET, VENDORS, AND SERVICE PROVIDERS

Evolution of SANs' Market

SAN technologies allow existing enterprises to effectively manage more transactions, customers, suppliers, and services. Company operations are significantly improved by providing continuous high availability through uninterruptible access to data, increasing scalability through multiple-channel data transmission, and reducing the network and server's CPU overhead. Additional opportunities for the IT enterprises are also associated with the Internet, which allows them to increase the volume of data and rates of their transmission. According to the International Data Corporation (IDC), since the mid-1990s, the number of users of e-commerce services has increased exponentially up to several hundred millions. SANs have revolutionized the IT enterprise's infrastructure and improved its e-business applications including e-commerce, e-mail, online transaction processing, data replication, and enterprise database management. Global continuous delivery of multimedia secured information has become the main service of modern e-business enterprises.

By adding networking and intelligence features to data storage, fibre channel SAN switches enable the solution of several challenging e-business storage problems, such as linking high-performance workstation clusters, connecting high-performance tape storage on disk farms, giving server farms a high-speed data-transmission pipe, clustering disk farms, and linking Ethernet, FDDI, ATM, and token ring LANs to the backbone. Intelligent SAN systems allow improving enterprise performance significantly, decreasing latency, supporting direct access to the storage shared by multiple servers, reducing network traffic on the front-end network, and removing storage management tasks from servers.

In December 2000, the IDC Corp. estimated that worldwide disk storage systems sales were about $31.7 billion in 2000. Networked-based systems with SAN installations represented 20% of the 2000 revenues. Trends indicate that both SAN and NAS implementations are accelerating, as SANs have grown 70% year over year. This trend is expected to continue during the next 5 years. Networked storage on the whole experienced a 43% capital asset growth rate during the same period, with SANs growing at 33%.

SAN Vendors and Service Providers

Table 2 represents a list of SAN vendors, storage-networking service providers, and their products. The complete list of SAN deployment companies can be fount in (Vacca, 2002, pp. 495–497) and on the Network Buyers Guide Web site (Network Buyers Guide, 2003).

The main providers of SAN solutions are EMC Corporation (about 40% of the market), Compaq Computer Corporation (13%), Sun Systems, Inc. (11%), IBM Corporation (10%), Hewlett-Packard Company (7%), Dell Computer (3%), Hitachi Data Systems (2%), Brocade Communication Systems, Inc., SANgate, TrueSAN Networks, Inc., and XIOTech Corp.

Table 2 SAN Vendors and Service

Company	Product Type	Source
ADVA Optical Storage Networking Technologies	DWDM system	(ADVA, 2003)
ATTO Technology, Inc.	Fibre Channel hub	(ATTO, 2003)
Brocade Communication Systems, Inc.	Fabric switch	(Brocade, 2003)
Computer Network Technology	Storage router	(CNT, 2003)
Compaq	Disaster-tolerant SAN solutions	(Compaq, 2003)
Crossroads Systems, Inc	Modular storage router	(Crossroads, 2003)
Cutting Edge	Clustered failover	(Cutting Edge, 2003)
Dell	Enterprise storage solutions	(Dell, 2003)
EMC Corporation	Networked storage solutions	(EMC, 2003)
Emulex Corporation	VI/IP PCI host bus adapter	(Emulex, 2003)
Hewlett-Packard	SAN management tools	(Hewlett-Packard, 2003)
IBM	Enterprise storage server	(IBM, 2003)
LSI Logic Storage Systems, Inc.	Storage manager software	(LSI, 2003)
McData	Enterprise FC management tools	(McData, 2003)
Media Integration	Fibre channel SAN; open storage	(Media Integration, 2003)
Nishan Systems	IP storage switches	(Nishan Systems, 2003)
Procom	Storage systems	(Procom, 2003)
Storage Computer Corp.	Storage systems	(Storage Computer, 2003)
StorageTek	Enterprise fibre channel switch	(StorageTek, 2003)
SUN Microsystems	Storage systems	(Sun Microsystems, 2003)
U.S. Design Corporation	RAID storage hardware	(US Design, 2003)

CONCLUSION

SANs, networked high-speed infrastructures, enable e-business enterprises to improve significantly their 24×7 continuous scalable services. They have become a critical part of the enterprise network infrastructure. The above-considered technologies and effective SAN solutions allow companies to shift their focus from numerous IT infrastructure problems to the successful performance of their businesses and services.

GLOSSARY

CIFS Common Internet file system, also known as the Microsoft server message block protocol; a network file system access protocol that is primarily used by Windows clients to communicate file access requests to Windows servers.

CIM Common information model; an object-oriented description of entities and relationships in the enterprise management environment.

DWDM Dense wavelength division multiplexing; a method that allows more wavelengths to use the same fiber.

FC-AL Fibre channel-arbitrated loop transport protocol.

FCSW Fibre channel switch.

FC-VIA Fibre channel–virtual interface architecture.

FSPF Fabric shortest path first; a routing protocol used by fibre channel switches.

IPoFC Internet protocol over fibre channel.

iSCSI Internet small computer systems interface.

JBOD Just a bunch of disks; a term for a collection of disks configured as an arbitrated loop segment in a single chassis.

NAS Network attached storage.

RAID Redundant arrays of inexpensive disks; a technology for managing multiple disks.

SAN Storage area network.

SAS SAN attached storage.

SoIP Storage over IP; a storage technology developed by the Nishan Systems Corporation.

SSA Serial storage architecture.

VI Virtual interface architecture; a midlayer protocol specification.

CROSS REFERENCES

See *Conducted Communications Media; Standards and Protocols in Data Communications; TCP/IP Suite.*

REFERENCES

ADVA optical storage networking technologies (2003). Retrieved March 13, 2003, from http://www.san.com/

American National Standards Institute (ANSI) (2003). Retrieved March 13, 2003, from http://www.ansi.org/

ATTO SAN solutions (2003). Retrieved March 13, 2003, from http://www.attotech.com/sans.html

Barker, R., & Massiglia, P. (2001). *Storage networking essentials: A complete guide to understanding & implementing SANs.* New York: Wiley.

Beauchamp, C., Judd J., & Kuo, B. (2002). *Building SANs with Brocade fiber channel fabric switches.* Rockland, MA: Syngress Publishing, Inc.

Brocade Communication Systems, Inc. (see the SAN Info Center) (2003). Retrieved March 13, 2003, from http://www.brocade.com/

Clark, T. (1999). *Designing storage area networks: A practical reference for implementing fiber channel SANs*. Boston: Addison–Wesley.

Clark, T. (2002). *IP SANS: An introduction to iSCSI, iFCP, and FCIP protocols for storage area networks*. Boston: Addison–Wesley.

CNT (Computer Network Technology Corp.) storage over IP solutions (2003). Retrieved March 13, 2003, from http://www.cnt.com/

Compaq, Inc. (2003). Retrieved March 13, 2003, from http://www.compaq.com/

Crossroads Systems, Inc. (2003). Retrieved March 13, 2003, from http://www.crossroads.com/about/

Cutting Edge, Inc. (2003). Retrieved March 13, 2003, from http://www.cuttedge.com/

Dell, Inc. (2003). Retrieved March 13, 2003, from http://www.dell.com/

Distributed Management Task Force (DMTF) (2003). Retrieved March 13, 2003, from http://www.dmtf.org/

EMC Corporation (2003). Retrieved March 13, 2003, from http://www.emc.com/

Emulex Corporation (2003). Retrieved March 13, 2003, from http://www.emulex.com/

Farley, M. (2001). *Building storage networks* (2nd ed.). New York: Osborne/McGraw–Hill.

FCIA (Fiber Channel Industry Association) (2003). Retrieved March 13, 2003, from http://www.fibrechannel.org/

Fibre Alliance (2003). Retrieved March 13, 2003, from http://www.fibrealliance.org/

Hammond-Doel, T. (2001). *2 Gb/s fiber channel SANs*. Vixel Corporation. Retrieved March 13, 2003, from http://www.vixel.com/9000_docs/9000_wp.pdf

Hewlett-Packard, Inc. (2003). Retrieved March 13, 2003, from http://www.hp.com/

IBM Enterprise SAN Solutions (2003). Retrieved March 13, 2003, from http://www.storage.ibm.com/ibmsan/

IEEE SSSWG (Storage Systems Standards Work Group) (2003). Retrieved March 13, 2003, from http://www.ssswg.org/

InfiniBand Trade Association (2003). Retrieved March 13, 2003, from http://www.infinibandta.org/home/

InfraStor Technology Corp. (2001). *SAN vs NAS*. Retrieved March 13, 2003, from http://www.infrastor.com/tech/sanvsnas.htm

InfraStor Technology Corp. (2002). *Introduction to SAN*. Retrieved March 13, 2003, from http://www.infrastor.com/tech/SANTechIntro.htm

Intel InfiniBand Architecture (2003). Retrieved March 13, 2003, from http://www.intel.com/technology/infiniband/

Internet Engineering Task Force (IETF) (2003). Retrieved March 13, 2003, from http://www.ietf.org/

Jini (2003). Retrieved March 13, 2003, from http://www.jini.org/

Jiro (2002). Retrieved September 12, 2002, from http://www.sun.com/jiro/

LSI Logic Storage Systems, Inc. (2003). Retrieved March 13, 2003, from http://www.lsilogicstorage.com/

McData, Inc. (2003). Retrieved March 13, 2003, from http://www.mcdata.com/

Media Integration, Inc. (2003). Retrieved March 13, 2003, from http://www.mediainc.com/

National Storage Industry Consortium (2003). Retrieved March 13, 2003, from http://www.nsic.org/

Network Buyers Guide (2003). Retrieved March 13, 2003, from http://networkbuyersguide.com/

Nishan Systems (2003). Retrieved March 13, 2003, from http://www.nishansystems.com/

Ottem, E. (2001). *Third generation SANs: Open 2 Gb fabric* (White Paper WP-2G0801). Gadzoox Networks, Inc. Retrieved March 13, 2003, from http://www.infrastor.com/downloads/2Gb_whitepaper.pdf

Peterson, M. (1998). *Storage area networking*. Santa Barbara, CA: Strategic Research Corp. Retrieved March 13, 2003, from http://www.sresearch.com/wp_9801.htm

Procom Technology Corporation (2003). Retrieved March 13, 2003, from http://www.procom.com/

Sachdev, P., & Arunkundram, R. S. (2002). *Using storage area networks*. Special edition. Indianapolis, IN: Que.

Sheldon, T. (2001). *McGraw–Hill encyclopedia of networking & telecommunications*. New York: McGraw–Hill.

SNIA (Storage Networking Industry Association) (2003). Retrieved March 13, 2003, from http://www.snia.org/

Storage Computer Corporation (2003). Retrieved March 13, 2003, from http://www.storage.com/

StorageTek, Inc. (2003). Retrieved March 13, 2003, from http://www.storagetek.com/

Sun Microsystems (2003). *Data storage solutions*. Retrieved March 13, 2003, from http://www.sun.com/storage/

Thornburgh, R. H., & Schoenborn, B. J. (2001). *Storage area networks: Designing and implementing a mass storage system*. Upper Saddle River, NJ: Prentice Hall PTR.

Toigo, J. W. (2001). *The Holy Grail of data storage management*. Upper Saddle River, NJ: Prentice Hall PTR.

U.S. Design Corporation, Inc. (2003). Retrieved March 13, 2003, from http://www.usdesign.com/

Vacca, J. (2002). *The essential guide to storage area networks*. Upper Saddle River, NJ: Prentice Hall.

Strategic Alliances

Patricia Adams, *Education Resources*

STRATEGIC ALLIANCES IN E-COMMERCE

Strategic alliances are separating competition. The winners are creating "best of the best" branding by forming teams of people and "completing" companies from past "competing" companies. Today, more than ever, most businesses don't have the desire, capacity, or resources to grow without the help of other aligned partners. Former paradigms of business development are being challenged daily as new ideas, new companies, and innovation provide the leadership for others to follow.

STRATEGIC ALLIANCES—THE WAY TO SUCCESS IN BUSINESS

Trading Partner Relationships

A "win/lose" mindset that exists in many trading partner relationships is a dynamic that has kept value chains from achieving a desirable performance. As stated in the Voluntary Interindustry Commerce Standards (VICS) Association for the Collaborative Planning, Forecasting, and Replenishment (CPFR) Committee, a VICS committee made up of retailers, manufacturers, and solution providers, "Under the CPFR approach, companies have to look across value-chain processes to see where their information or competencies can help the value chain and thus benefit the end consumer and the value-chain partners. This outlook acknowledges that nobody wins until the *consumer is satisfied*. One of the keys to the success of CPFR, in other words, is a changeover from 'win/lose' to 'win/win' value-chain relationships" (*Corporate Values*, n.d.).

On the CPFR Web site, http://www.cpfr.org/GuidingPrinciples.html, are the guiding principles developed for CPFR out of the JMI and VMI best practices: "In the course of this discussion, we examine traditional aggregate forecasting and replenishment as well as vendor-managed inventory (VMI), and jointly managed inventory (JMI)" (*Current Process State*, n.d.).

The trading partner framework and operating process focus on consumers and are oriented toward the value chain success.

Trading partners manage the development of a single shared forecast of consumer demand that drives planning across the value chain.

Trading partners jointly commit to the shared forecast through risk sharing in the removal of supply process constraints.

Various trading partnerships exist today for unique enhancements to the relationship. A vendor exists to sell something, a subcontractor to work for someone, a tactical partner, possibly to oversee the partnership, and a strategic partner, possibly integrated for several reasons of interdependence that could also include a joint venture partner, developed through an acquisition or a merger.

Partnerships vs. Strategic Alliances

A "partnership" in business typically suggests a legal agreement of participating in a venture of some sort for a very specific outcome. Real estate partnerships, for instance, might be formed to accumulate funds to buy real estate for the partnership. In *Webster's New Collegiate Dictionary,* partnership is defined as a legal relationship existing between two or more persons contractually associated as joint principals in a business. A partnership may or may not be represented by a written contract but could still be formed legally. "Strategic" in the same dictionary is defined as "of great importance within an integrated whole or to a planned effect." "Alliance" is defined as "an association to further the common interests of the members."

Strategic alliances is a term that has been used most recently as it has related to commerce through the use of the Internet. On Cisco's corporate Web site page is an explanation of their philosophy: "Our strategic alliances form a customer-centric, total solution approach to solving problems, exploiting business opportunity, and creating sustainable competitive advantage for our customers" (Cisco Systems, 2002a).

At Cisco, **strategic alliances** engage in **impact partnering** that is differentiated by eight **unique but integrated** characteristics:

- Industry Leaders,
- Multiple Touch Points Across Both Companies,
- Cross Value Chain Impact,
- Joint Solution and/or Technology to Address Customer Needs,
- Long Term Investment of Resources and IP,
- Anticipate Competitive Threats,
- Create New Global Markets and IP, and
- Weave Multiple Partners Together to Target New Markets (Cisco Systems, 2002b).

Although partnership and strategic alliance can be used interchangeably in some circumstances, the formation of a strategic alliance seems to insinuate a relationship of interdependence, so important that the entity cannot exist without the sum of all of those involved. As stated in the *ASCI Journal of Management* in 1992, strategic alliances have brought about a new dimension to a globalizing economy. They have led as well as are being led by various shifts in the global markets (Asma, 1992).

Partnerships as Strategic Alliances Development Prior to E-commerce

Creating partnerships, legal or otherwise, is not a new concept in business. In ancient history, bartering was a common form of existence because it provided a way to obtain products or services through an exchange. Typically, merchants focusing on developing a certain skill to create a product or service depended on others for products or services that they had no skill or abilities to produce. This type of business arrangement provided a certain amount of stability and peace to a village.

Franchising concepts have been another form of creating partnerships. Franchising is at least 150 years old. One early example resulted in the characteristic look of historic hotels (bars) in New South Wales, with franchising agreements between hotels and breweries. An American example was the telegraph system operated by various railroad companies but controlled by Western Union.

Multilevel marketing (or network marketing), such as used in the Amway business, is another partnering business model. Typically, independent business owners are associated with the company in a contractor-like relationship. A percentage of the profits from the sale of goods is distributed between the relationships.

Strategic Alliances in E-commerce

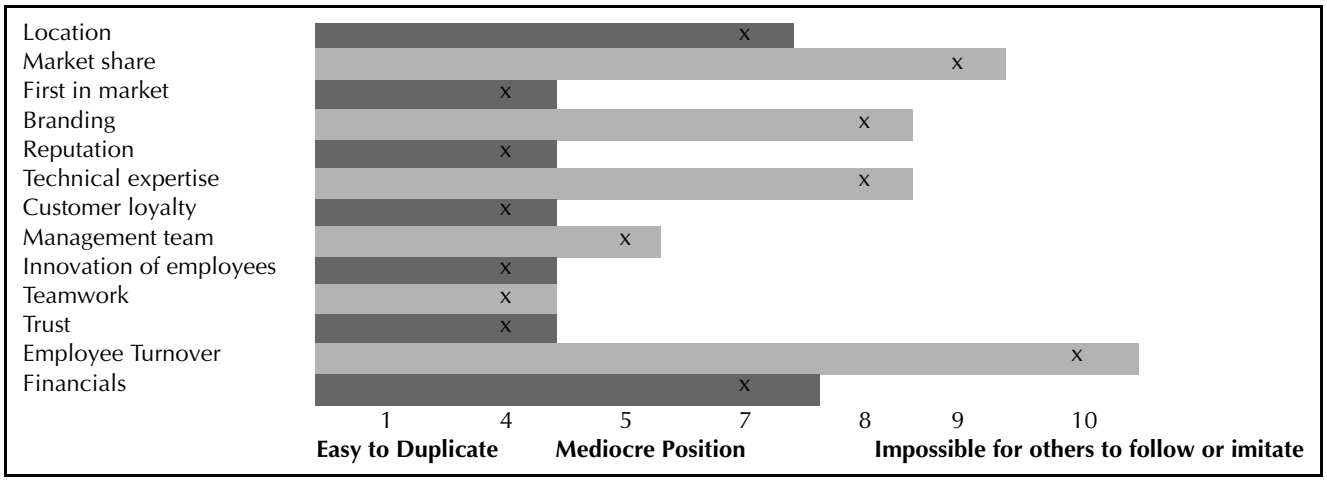
As the Internet has grown so exponentially in use, so has commerce activity over the Web. Businesses have been forced to create specific partnerships as a means of working through the complexities of producing, selling, and distributing products over the Web. At the same time, they have been required to create a strategy for a totally new way of connecting to the customer. The broadened exposure to the market and the speed requirement induced by that exposure coupled with the speed expectations of the customers has furthered the need to find and develop appropriate strategies. In contrast to earlier times when skills were limited, the use of the Internet has added so much complexity that it has once again forced business owners to focus on doing one thing really well.

The speeding up of finding, ordering, paying, and demanding for delivery has caused a need for companies to focus on their strengths, and at the same time, eliminate the processes within their company that others can do better. As downsizing and financial burdens have increased in recent years, businesses are more focused on "outsourcing" their needs as a means of balancing the budget. Focusing on the highest and best use of company resources has required development of synergism with other companies.

Identifying and Developing Core Competency

The processes identified to develop a clear understanding of a firm's core competency must begin with the simple question: What do we do better than anyone else? Some of the most important core competencies are listed in Table 1. As a company begins a process to consider the advantage they have over competitors, they also need to assess how important, or not, those characteristics might be to the overall strategies. Table 1, for instance, gives the opinion that while a company might be the first company to come into the market with a particular invention, they might not be the company that ultimately makes a success of that invention. The table estimates a value of 3 out of a possible 10 for how the importance of being first to market might be compared to other indications of a company's ability to be successful. The intent of the graph is to show the likelihood of a competitor being able to surpass a firm's particular competency. This graph, obviously, values employee stability as one of the most important factors in

Table 1 Core Competency Indicators Grid



the success of a company. One might conclude that it is more important to spend time and dollars on employee development and general satisfaction than in branding if budget is at issue.

Comparing Strategic Alliances for Different Businesses

As companies have focused on their core business mission and have looked to others to complete their desired business model, interesting relationships have begun to emerge. As computer manufacturers searched for ways to differentiate themselves from other similar firms, they looked for software other companies had developed for use as operating systems many years prior to the Internet.

As Gateway Computers strived for a market share greater than that held by Dell Computers and others during the 1990s, they looked for new ways to expand. They began to create strategic alliances with software companies to develop and produce programs that Gateway could include on their computer desktops. They created hundreds of these strategic alliances as a means of outperforming their competition. During the same time, several small businesses beginning to establish an Internet presence were trying to identify potential partners that might provide additional market exposure. Many of these smaller companies did not have the resources to develop marketing and human resource services "in-house." The blending of these companies' strengths created opportunities previously unattainable.

In another example, during 1999, CollegeCapital, an Internet college preparatory and scholarship database company developed a strategic alliance with American Express. They provided resources to American Express families as American Express provided exposure to their enormous market. Both companies benefited from each other's third-party recommendation as they both had developed reputations in their specific markets of strength and integrity.

Strategic alliances usually provide several equitable benefits to each party in addition to revenue sharing, including the use of each other's sales forces, executive leadership, and marketing ideas. Careful consideration should be focused toward an identifiable, specific outcome before the formation of any type of partnering association.

STRATEGIC ALLIANCE APPROACH TO ENHANCEMENT OF DEPARTMENTS WITHIN A COMPANY

Through corporate reorganization and downsizing, many executives have been left without jobs or reengineered to jobs that pay less. At the same time, small companies have been unable to afford human resource departments or assistants that at one time seemed imperative. As executives became proficient with the use of computers and technology, the need for many support jobs once considered necessary was lessened or eliminated. Circumstances like these have drastically changed the market in general and have also lead to a boom of support-related services and small business growth. Previously, where one secretary worked within a corporation for one executive, many former secretaries have now started secretarial companies, supporting several companies with secretarial outsourcing. Through using these types of strategic alliances, companies are able to downsize or add support almost overnight with very little disruption to ongoing business. Professional corporations previously were most known to be attorneys, accountants, and physicians. Since the Internet boom, with the complexity and speed of technology, small businesses as professional corporations have formed to provide outsourcing for housekeepers, personal chefs, Web site developers, software developers, secretaries, human resource directors, creative designers, and consultants of all types. The list includes companies focused on creating strategic alliances.

Enhancing Executive Management through the Use of Strategic Partners

Accounting scandals over compensation of senior executives, corporate bankruptcy, and misuse of company

assets uncovered during 2002 created an unprecedented economic turmoil. Executives, accustomed to comfortable corporate salaries, suddenly found themselves without jobs. Lack of corporate opportunities caused several senior executives to develop creativity in marketing themselves as strategic partnering consultants. This arrangement as the newest form of a management team has lead to an attractive alternative and creation of a win/win circumstance for companies and individuals. Talented executives have been able to work for several companies at once, using the best of their own resources and have been able to choose their schedules, work habits, and companies for whom they want to work. For executives who chose to work through consulting companies, many were able to make more money and experienced greater flexibility in how and when they worked. Larger companies have benefited by being able to hire extremely talented individuals, at a fraction of the cost of a full-time executive employee, participating in a sort of "executive" time-share with other companies. Companies can bring in the best people as a brain "tank" for less money and the consultants are glad to be there. Establishing an environment surrounded by excited and appreciative new individuals creates enthusiasm as well, throughout the organization and for long-time employees. Charrettes can be formed in a short amount of time using the most cutting-edge information and executives, and can lead a company out of a tremendous upheaval toward new successes. This paradigm shift being experienced by many companies today seems to be creating the new ideal business model. Although this type of business arrangement is a fairly recent one, these partnerships seem to be established for the long-term gain.

Enhancing Marketing through the Use of Strategic Partners

Marketing, branding, advertising, and public relations have followed the rest of the business world in becoming increasingly more complex and competitive. Most large corporations previously created marketing solutions through their own employees, sometimes as a human resources department function or within their marketing group. The Internet and sophisticated systems of marketing have created a need for a focused understanding and knowledge of marketing as a complete system of tools and branding that will constantly update and change according to rapid changes in market conditions. Public relations firms have had to reorganize drastically in recent years to be able to handle the complexity of maintaining market share in a global market, using the Internet as well as all other forms of media. The statistics and effects of certain marketing techniques seem to reinvent themselves daily. The interests and desires of customers tire easily as massive groups use the Internet. Web sites experiencing a great amount of traffic will many times change their look once a day to keep customer interest, and several companies change the front page of their Web site upon each computer visit.

Creating strategic alliance partnerships between marketing companies and businesses as a means of revenue sharing has been a boost for many marketing companies as well as for their host businesses. Traditionally, marketing companies billed for work performed and were left out of the windfall of financial success that came from growing start-up companies. As companies develop best of the best branding, some have decided to form a revenue-sharing process with marketing companies as an incentive to help build market share. The company needing marketing benefits from an alignment with a great marketing company at an early discounted rate while it is small and trying to build growth. As market share increases, the marketing company's revenue sharing increases exponentially, a benefit that previously was not an option for payment. This type of arrangement could build residual future incomes as well for the marketing company, depending on the particular contract agreement.

Enhancing Supply Chains through the Use of Strategic Partners

In a news release of October 1997 on Boeing's Corporate Web site, Boeing created a revolutionary partnership to obtain a large U.S. Army contract. By forming a partnership with Apache, Lockheed Martin, and Army Depot, a repair work company, they were able to combine cost saving "support costs while bolstering the Apache's readiness capabilities" (Boeing, 1997).

In 2000, Gateway Computers and CollegeCapital created a partnership that would expand the market of each company while providing resources to each of their customers. Gateway, as a means of creating distinction from competitors, wanted to create software solutions as an added incentive to purchase Gateway Computers. By adding CollegeCapital's membership to the sale of a computer to a student's family, Gateway was able to add greater value to the computer purchase. CollegeCapital developed mass purchase discounts for individual student memberships in exchange. The exposure for CollegeCapital in Gateway Country Stores created a dynamic growth to their market. By combining marketing campaigns, both companies benefited from each other's endorsements to their perspective customers. CollegeCapital was able to provide computer discounts to students through special pricing incentives from Gateway, and the customers were able to obtain better pricing, a better overall product, and ease of delivery. Both companies benefited from obtaining a greater percentage of sales and market share and were able to create cost savings in many support areas where only one company's resources were needed. Teams of employees from each company were shared and excellent ideas formed by the diversity that was created in bringing the expertise of employees together from completely separate companies and corporate cultures.

Enhancing Delivery through the Use of Strategic Partners

There was a time when the doctor would come to the home of an ill patient, prescribe medication, and then deliver it to the patient. Of course, the doctor had a limited choice of medications contained in the little black bag, the downside of a simple solution.

Today with the introduction of the Internet and e-commerce, delivery has become one of the most demanding and difficult management problems. As customers simultaneously connect to a merchant's site, the computer takes the order, captures the payment, and sends notice to have the order filled. There is almost an endless capacity for the number of orders that can be obtained at the same time. The need for human interaction becomes an essential element to the automated ordering procedure in many cases. If an order was created inaccurately, a customer most likely desires to call someone. The order then needs to be fulfilled and shipped. This slowing of the overall procedure has created new businesses that focus on specific aspects of these processes. Huge call centers having 24-hour access have formed almost overnight. Fulfillment houses, warehouses full of several different products from several different companies, have been created. Product delivery is a system that has become so competitive, complicated, and focused that there are companies specializing solely in overnight delivering. In some cities, like New York, delivery of office products to waiting businesses involves bicycle delivery because of traffic congestion. It seems in many circumstances that there has been a simplicity reversion, of sort, to the "old days" in order to deliver and obtain desired goods immediately.

Creating strategic alliances from necessary delivery partners has demanded unique approaches. If a company's service is to guarantee a specific delivery, they must have a dependable partner. Revenue sharing has been developed in some cases, to create a specific strategy to ensure dependable delivery. Delivery companies benefit by creating a built-in customer base and are able to advertise that they are partnered with a company known for excellent customer service. An example of this type of an alliance was referenced in the book *Getting Partnering Right*. The authors wrote, "FedEx has partnered with Intel to take over part of Intel's logistics. As a result, guaranteed delivery time has improved from four business days to three—and delivery errors have been substantially reduced" (Rackham, Friedman, & Ruff, 1996 p.13).

During the summer of 2002, one physician in Los Angeles, CA, decided to take his office to his patients. An improvement over the "black bag," by using today's technology and several partner teams, patients are able to have access to very sophisticated procedures while at home.

STEPS TO FINDING THE RIGHT STRATEGIC ALLIANCE PARTNERS

It is obvious in the complexity of e-commerce that one company cannot "do it all." Processes, regulations, ideas, and human potential have grown to require an enormous amount of expertise and information. Daily, the news carries stories of large corporations, past "do it all" companies, now in trouble or already bankrupt. They missed the mark. The larger some companies have become, the more difficult it has become to readjust to competition at the speed required to stay on top of the market.

How do companies create partnerships that will be mutually beneficial? How do companies even know that they need partners? What steps can be taken for older, established companies needing to be retooled? How can partnerships be established when a company is just beginning to form as a new e-commerce venture? Table 2 provides a checklist for determining strategic alliances.

CASE STUDY

During the past several decades of business, as a company grew, it usually expanded the various departments needed to fulfill its business model. During the 1970s, for example, a developer in Arizona started a home-building company by obtaining land and partnering with a contractor who built houses. Several subcontractors were hired to perform specific tasks: a large equipment operator, a cement contractor, a framer, a roofer, and so on. The profits of the development were split between the company owning the land and the general home-building contractor. As the company grew during a 20-year period, it began to form various departments within its own organization, eventually buying out the original home-building contracting partner. The development company grew to over 1,200 employees in more than 100 intercompany departments. The developer owned the water company, the landscaping company, the large equipment operations, the cable television, and even the mortgage company. As e-commerce began, customers were able to see home-building options and several home developments advertised on the Web. They compared building plans with other developers across the country, simultaneously. Mortgage rates were compared against Web-enabled mortgage loan applications, rates, and funding. Buyers could shop for trees and landscaping packages advertised on the Internet that once were sold as part of the package with the house. Cable television was being compared to other sources for television reception delivery. This explosive change caused the developer to reengineer the entire company. Layoffs occurred in many departments as the developer had to refocus on exactly the things that his company could do better than the competition. After the company was restructured, the obvious competitive advantage was their ability to build beautiful homes on unique land lots. Using years of accumulated customer information, knowledge as to what the buyers wanted and how decisions were made, they began to focus their marketing efforts through the use of strategic alliance marketing partners, instead of through additional employees. Partnered marketing companies focused on staying ahead of marketing trends, not only as related to home building but also other trends that might effect a home purchase. The development company downsized, and closed many departments, replacing them with consultants and strategic alliance partners.

The development company was able to view the enormous change in the market and react quickly. Unfortunately, many large companies have been lost because of the inability of key executives to see the big picture driven by e-commerce changes. If embracing innovation was not a key ingredient of a company during the past 20 years, the company has probably failed during the past 5.

The steps taken during the reorganization of the development company began with a focus on their core competencies. All key executive and management personnel

Table 2 Checklist for Determining Strategic Alliances

Internal
Assess the willingness to change by each employee.
Is management crisis orientated or proactive?
Are all processes and employees documented for accountability?
Is customer service a priority and does it start with internal customers, employees, vendors, etc.?
What current systems need to be revamped?
What is the result of the analysis of the firm's core competency indicators?
Current market conditions
What are the priorities to the current market?
What are the top core competencies of the firm in relation to the market?
Does the market demand growth?
Can the present company culture handle rapid growth/change?
Growth potential
Identify any examples of possible growth/expansion.
Identify positives/negatives to growth including cost vs/profit, using 80/20 rules.
Identify unknowns regarding results of possible expansion.
Next step: Think tanks
Develop several think tank and focus groups of trusted advisors, including good and difficult customers, employees, and potential partners to further discuss outcome of this checklist.
Implementation
Begin to identify a list of potential partners or strategic alliances that could see value in the present company's core competencies and develop a process to explore the likelihood of the partnership.
Specific desired outcome
As potential partnerships are identified, develop clear specific desired outcomes compelling to each entity.

were sequestered for several days to come up with ideas that would create a path for long-term survival, using an outside consulting group. A "think tank" approach was used to inspire innovation and creativity within the group. Through a series of questions and answers over several days, the company was able to bring into focus the things that they could do better than others and the things that they must do if they were to stay competitive. Recommendations for immediate and long-term planning were created. Hiring and partnering outside of the company was the new "norm" and getting rid of partially productive departments saved the company's core assets.

Timelines were developed according to the seriousness of the impact of the change. Cost savings and the ability to compete within the new e-commerce environment were of primary importance.

Let us review the steps taken by the development company in reengineering corporate structure to compete in an e-commerce environment:

Market changes were noticed and calibrated.

Key decision-makers met in a sequestered environment to participate in a "think tank" to formulate plans for changes.

"Highest and best use" was developed for all departments and key executives.

The company mission was changed to give better focus.

Weaknesses of the company were identified as possible areas for strategic partnering needs and as a means of identifying unnecessary use of company assets.

A new business plan to establish the new mission was developed.

Restructuring occurred to eliminate waste and unnecessary departments and employees.

Goals were established for "If we had strategic partners, what could be accomplished?"

Assignments were made with appropriate personnel to set meetings with potential partners.

Compatibility in key areas were thoroughly discussed between stakeholders of the companies involved, which included

- Corporate missions;
- Corporate culture: attitudes, dress, work structure philosophies;
- Time in business;
- Customer service attitudes and procedures;
- Marketing budget, present and future;
- Reputation with vendors, subcontractors, employees and customers; and
- Employee pay and benefits;
- Government and legal compliance.

Procedures to establish appropriate strategic alliance with long-term contracts as the goal were developed.

CREATING A LARGER COMPANY MISSION

There have been many articles and books written on how to create a company mission, but with the introduction of e-commerce, new questions that once were not important have had to be considered. Because the Web introduces

the possibility of worldwide exposure, a larger look at potential should be assessed from the very beginning. Any company mission should create a long-term focus with a broad use of the mission.

It is obvious, more than ever, that individuals and companies must know exactly what they are best at, or their highest and best use. Once an individual or company understands their truly unique talents and abilities, they can begin to encourage and grow those talents, through more specific education and experience. Once the core competencies are recognized and defined, focus and direction to develop those competencies can occur. Money is not wasted trying to grow something that at best will be mediocre in the competitive arena. Today, because competition is a global concern, there is no need to encourage a business, a department, or an individual to be or do something that cannot be maintained in excellence. With the added complexities and extreme competition in today's business environment, a company must do what it does better than anyone else. To do otherwise is a prelude to failure.

IDENTIFYING THE PART YOUR BUSINESS PLAYS

As a business answers, "What do we do better than anyone else?" the list becomes shorter and shorter. The next question then becomes "What are we doing now that we could get someone else to do better?" More often than not, it is a perceived direct competitor that has an ability that could be very beneficial if resources were combined. Competitors, when partnered, typically can become complimentary, each company focusing on what they do best and sharing in the support teams of employees. This thinking, although seeming common sense, requires individuals, as well as companies, to look at what drives them. If individuals are untrusting and in control of a company, they will be unable to develop a company based on respectful relationships and open communication. These attributes are imperative when creating strategic alliance partnerships. Unfortunately, many CEOs tend to hire persons with similar traits, and when they are of a suspicious nature, they unknowingly tend to surround themselves with similar individuals. This practice within a company, large or small, is one to avoid. Ultimately, strategic alliances will only work if integrity is woven throughout the organizations involved. So the rules for building strategic alliances are

Rule 1: Find partners that your company wants to be around, those your company can respect and work with, as a strategic alliance partner, whatever reputation they have, will be attributed to your company.

Rule 2: Find partners who have an established reputation of market strength, stability, and history in their focus.

As the mission is being developed and focus is identified, the mission will be expanded to encompass growth of that core competency and greater market share of that focus. The focus may change to becoming more of a business-to-business partner with strategic alliances than marketing directly to the public.

As CollegeCapital's business model continued to evolve from its inception, it became obvious that its best use of developing and providing the largest secured database of scholarships would facilitate schools helping students. In the original formation of its business plan, CollegeCapital was to help students directly, one at a time through connecting with them over the Internet. The model had to expand in order for CollegeCapital to stay focused on its core business, which was to maintain and sell information contained in their database of scholarships. Schools were in the business of connecting to and providing service for students. CollegeCapital decided to find partners with those companies that needed what they had. It had not developed a business based on understanding how students connect to information regarding college preparatory and scholarships. Schools typically provided that connection. One of the largest partnerships occurred for CollegeCapital from forming a strategic alliance with Cox Communication and Learning Station, companies connecting schools through an ASP design, delivering curriculum and information directly to the classroom. The strategic alliance brought CollegeCapital back to focus on its core business, that of developing the scholarship and college information, bringing value to those able to deliver it. The companies were able to share in revenue, marketing, and expertise using each other's employees and partners. Mutual trust and communication became the necessary component of the relationship. Cox Communication and Learning Station trusted that CollegeCapital would stay focused on the creation of the information that they needed to deliver. CollegeCapital trusted that Cox Communication and Learning Station would continue to focus on ensuring that the schools had a connection to the information. As the alliance grew, just as in most relationships, so did other opportunities. The project for Arizona schools was the largest technology project with the most partnerships ever attempted.

Change is usually upsetting and although most companies and individuals have felt greater stress, aggravation, and complication with the many changes of recent years, the formation of strategic alliance partners can actually simplify an organization into a more focused effort. Instead of trying to be good at many different businesses within one corporate structure, a business has an opportunity to use very focused efforts to expand what it does best. This focused effort filters throughout a corporation to individual effort. When individuals focus on developing their own core competencies, usually what they enjoy doing and are uniquely good at, they add much greater value to their company. Companies are able to benefit from employees who master and use their greatest talents instead of trying to create a mediocre mix of abilities of several talents.

Identifying Strengths and Weaknesses

In 1994, Larry Wilson in his book, *Stop Selling Start Partnering*, quoted Alvin Toffler from his revolutionary best-seller, *Future Shock*, "No generation has witnessed so many simultaneous changes that are interrelated and of

a global nature.... On many, many fronts, we're in for a couple of decades of rattling, shaking and reorganization" (Wilson, 1994). In 1994, the author did not realize the complexities of the changes that were about to occur and the speed at which the market would continue to do so.

Just as a person takes tests to help identify their talents and personality type, and then embarks on a lifelong journey of self-improvement, businesses must conduct a checkup a minimum of at least once a year, especially now with market changes occurring so rapidly. There are experts in the field of organization identity and branding that all businesses should involve from time to time, at least to gain a different perspective from outside of the organization. Key decision-makers should meet to look at reorganizing or restructuring in order to meet the needs and changes of the current market. The need to constantly look at what you do, how you do it, and on what are you spending money and resources cannot be overemphasized. In the days of large corporate structures and jobs that lasted 20 or 30 years, there may have been an atmosphere that condoned complacency and waste. In today's market, however, "If you snooze, you lose." If one is not prepared to embrace change and constantly seek to become better, one will almost immediately begin to lose market share and marketability.

Who could have predicted that in one action American Express, a huge corporate structure, would be taken to its very core? The tragedy of September 11, 2001, and subsequent closure of airports left American Express travel customers stranded, trying to find a way home. Destroyed corporate headquarters and a damaged technology infrastructure was a catastrophe American Express and other companies could not have planned for. For weeks the travel business nearly stood still. Although American Express had grown considerably over many years of success, having been founded in 1850 and having added companies and complexities to their organization, they had maintained the core mission of what they felt that they did better than anyone else, to "provide world class customer service." When employees were unable to reach their managers and customers were not able to connect with travel guides, the essence of the company began to shine through the chaos. Employees from other departments, unrelated to travel customers, began to take frantic calls to help in rerouting. Employees throughout the organization began to pitch in to continue as much uninterrupted customer service as possible, even though facilities destruction was prolific. Phones were rerouted, technology was rerouted, and customer service continued even with the absence of strategic managers. The company was able to create change immediately to address extreme circumstances because individuals within the company had been trained to know and understand that at the core, American Express focused on maintaining outstanding customer service. The turnaround from losses caused by that fateful day was extraordinary. Stories of heroic efforts made by employees striving to continue to provide world-class service to their customers abounded.

In the fast-moving business of e-commerce, every individual within a company must internalize a company's mission. The simple truth of the Golden Rule becomes the dependable element particularly during a crisis. The story of American Express and its ability to recover from the catastrophe of 2001 is further evidence of the importance of being conservative and cautious when seeking like-minded, strategic alliance partnerships.

Create Compelling Reasons for Others to Seek Your Company as a Strategic Partner

Just as American Express continued to evolve its reputation of always maintaining outstanding customer service, so did those companies wanting to partner with them. While CollegeCapital knew that it had the best databases and information regarding college preparatory, no one else did. As they were reviewing the companies that might be willing to partner with them, American Express seemed to be a possible match. Identifying with a company that demanded outstanding customer service could have its challenges but would give an immense amount of credibility for a young company. CollegeCapital had developed information that could be used as an added benefit to the customers American Express already serviced. The student loan division of American Express, at that time, was a natural fit.

In order to create partnerships, individuals or corporations must obviously be desirable partners and others need to be able to recognize it. In forming a relationship with a well-known partner, a smaller company will have to establish evidence that they can bring a noticeable benefit to them. When a corporation is great at producing a certain product but is known for taking ruthless or less than ethical business advantage over other businesses, partnerships will be nearly impossible to maintain even if an initial agreement is reached. People should seek personal associations with those who provide strength rather than those who extract it and it is the same with strategic alliance partners. The best e-commerce partners are those that reflect positively on one another, creating greater substance, market share, and dynamic impact. The result, hopefully, is a partnership that is greater than the sum of each partner.

DEVELOPING AND MAINTAINING A STRONG RELATIONSHIP WITH STRATEGIC ALLIANCES

Comparing Benefits

Partnerships are born from a leader's vision of what can be achieved with the right circumstances and assets. A close look at competitors will typically reveal aspects of companies that if joined together could create a stronger market share for both companies or solve problems that are occurring. In e-commerce, for instance, it makes little sense to reinvent a software shopping cart system, even if the exact nature of what is desired is not readily available. A close evaluation needs to reveal whether it would be better to partner and share revenue with an existing shopping cart developer to create a new one, which would decrease future ongoing revenue, or to hire the work done and continue to support it.

A similar figure used in Rackham et al.'s *Getting Partnering Right* (1996, p. 12) represents the intersections

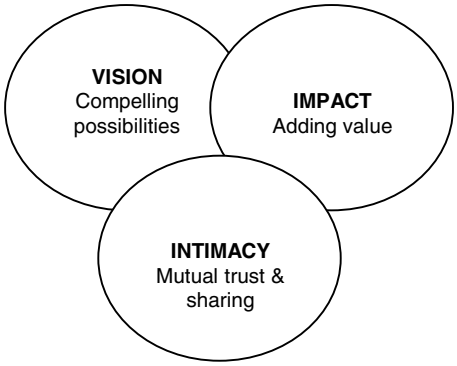


Figure 1: Spirit of partnering: trust, enthusiasm, and opportunities grow as the relationship intertwines. Source: Rackham et al. (1996). Reprinted with permission.

necessary for partnering to work and the level to which it can. As the amount of trust and sharing increases, so does the partnering effort and the mutual benefit. Attorneys usually develop necessary contracts but there is a degree of partnering that cannot be contracted for or ensured: it either is or is not a partnership based on mutual trust and benefit. Much like a marriage, it either works or it doesn't, depending on the attitudes behind the commitment. Partnering has a subtle ingredient that is nearly impossible to identify but supersedes all but intention. It could be called the "spirit" of the partnership, or the "spirit" behind it. As the components grow, dependency on the relationship also grows.

In order for a strategic alliance to work, everyone must have a stake, appreciation and respect for the proposed partnership. To the degree that greed, intolerance, difficult personalities, lack of sharing, and feelings of inequity enter the partnership, the relationship may erode. In a strategic alliance, there must be a compelling axiom: "We exist better because of the partnership." There is no reason to force a relationship or enter into a relationship that seems to be too difficult. Mutual respect for the various teams and employees must be shared by all of the employees involved in the alliance. Just as one negative employee can destroy the efforts, attitudes, and work of several within a company, one disgruntled employee can cause exponential damage to a more complicated structure of maintaining a strategic alliance. During the initial partnering effort, if it hasn't already occurred, it's time for a good "house cleaning," a cautious review of all employees' value, and values within the company. Anyone not adding specific, intrinsic value should be reevaluated with respect to his or her association with the organization.

Negotiations

In the world of creating strategic alliances for e-commerce, it is more important than ever to approach it from a standpoint of personal relationships. The old standby and still today as important a book, *How to Win Friends and Influence People* (Carnegie, 1936), references a conversation John D. Rockefeller had with Matthew C. Brush. Rockefeller said, "The ability to deal with people is as purchasable a commodity as sugar or coffee. And I will pay more for that ability than for any other under the sun."

The ability to negotiate the best partnership for all will depend largely on whether the principals want to spend time with each other. If they don't, no matter how important the deal, it probably won't work. Things have become so complex with the introduction of technology that we have had to resolve to the most simple of truths, the Golden Rule: "Do unto others that which you would have done unto you" (see http://www.fragrant.demon.co.uk/golden.html).

Before Closing the Deal: Identifying, and Resolving, Potential Problems with the Relationship

If it hasn't happened yet, get rid of the whiners in the company. Find employees willing to embrace change and be excited about working with a new strategic alliance. Communication, or lack of it, is always at the root of any problem in relationships. In developing e-commerce strategic partners, this cannot be overemphasized. A strong communication procedure must be developed to address any potential breakdown. Depending on the complexity of the partnership and the number of people and companies involved, it may make sense to use a Web-based communication or project management system. These systems only work when people use them, so tracking and requiring their use is imperative. The diversity of users and the number of work locations will determine the extent to which a Web-based communication project tracking system would be effective. Even when talks are begun between key executives, getting used to a Web-based system from the beginning can be very productive. Face-to-face contact and communication can never be underestimated, however. From time to time, it makes sense to bring the stakeholders together.

In the late 1990s, the State of Arizona wanted to create a way to provide a statewide connection for all students, teachers, and schools into one common system. In order to begin to address the complexities of such a mammoth project, a series of two-day "think-tanks" was implemented about six months apart. In the beginning, key decision-makers were invited to work at creating the solutions that would be needed. The "think-tank" events continue to be held for all of the partners involved in providing course content software on the system. During the events, they discussed communication problems and resolutions and the use of a Web-enabled project tracking system. The entire project partnered about 50 companies and political entities, some as strategic alliance partners, others as consultants and service providers to the project. The project had many stakeholders and requirements that needed to be addressed:

The State of Arizona would be assured that they had obtained discounted prices on all curriculum software, with an appropriate tracking system.

Statewide administrators would be able to get instant assessments, statistics, and communication throughout their district's schools.

Teachers would be able to have help in creating and augmenting their courses.

Children would have access to the most cutting-edge education content and information in a protective Internet environment, 24 hours a day, from any computer.

Parents would be able to communicate with their children's teachers to monitor their children's progress.

Cox Communications would be the first to develop the largest and most sophisticated ASP in the world through a partnership with Learning Station and others.

The main problem of any project of size is to create and maintain a system that allows and encourages, even demands, constant communication efforts. Communication works to eliminate barriers, crucial to all partnerships. Trust also becomes a critical ingredient in the beginning of any strategic partnership and continues to be an important issue throughout the relationship. As communication breaks down from time to time in all projects and partnerships, success, then, depends on proceeding with a collaborative effort of mutual respect. There must be a spirit of resolve: that one party is not trying to outdo the other and that everyone is important in maintaining a good communication and effort.

In the book *Getting Partnering Right*, the authors say, "In a complex partnership there will be so many uncertainties and unknowns that the path can't be charted with any precision . . . . The target that a shared vision is trying to hit is constantly moving" (Rackham et al., 1996 p. 118).

In the previously sited case of American Express and its disaster recovery, the spirit of the company ingrained in the employees ultimately solved the severe problems. A resolution projection for the issues created or for the possibilities of such a tragic event had never been designed. However, American Express as a company was able to weather the storm by the perseverance of employees and their commitment to the company mission.

The Contract

Once the expectation and vision for a partnership has been established, how does that evolve to a contract? In *Getting Partnering Right*, reference is made to a partnership document used by British contract specialist McGregor Cory, in which the opening simply states that it specifically disclaims any legal intent:

> This document is not written nor entered into as a formal or legal agreement but is only a definite expression and record of the purpose and intention of the parties, to which they each honorably pledge themselves. The document covers performance criteria and sets up agreed-upon communications channels. (Rackham et al., 1996 p. 20)

Perhaps business development is changing to be first dependent upon an honor system of what is right and just, and then building on character strengths for creating profits. Businesses and partnerships are recognizing that profit is a result of taking the necessary, honorable approaches.

There have been many articles written on how to create partnerships that avoid any antitrust issues. It should be stressed that all agreements avoid any potential conflict with regard to antitrust laws. It is important to obtain adequate counsel and information regarding the subject, and to make certain of all regulatory compliance issues. The U.S. judicial branch of the U.S. government's Web site has helpful suggestions and recommendations regarding this serious consideration.

The Internet, while giving greater control to the consumer, has eliminated many older traditional sales processes. The decision to buy today is based, as in the past, on price, integrity, dependability, service, and desire. The use of e-commerce brings instantaneous information and cross-referencing, giving consumers, instantly, an unparalleled amount of information and statistics they need in order to make decisions. The slightest issue, the smallest critical element, can make the difference in the sale. Dynamic, growth-oriented strategic alliance partnerships seem to be able to focus and deliver on details that customers search for and are demanding. These details should be part of the contract.

The contract should discuss all key talking points and consideration. The best contracts also require some resolution points for possible conflict that might arise into the future. Typically, the more complicated, longer contracts seem to be those that create the greatest aggravation to the forward purpose of the relationship. It is always advisable to obtain help from a contract attorney to finalize the document.

Win/Win? Or Not?

Is there any other way? In the past, several win/lose philosophies were invoked, perhaps not win/lose but at least win/more win. The philosophy of winning by causing others to lose has been rampant throughout corporate America. Employees were taught in many instances not to support each other and some felt that the only way to obtain recognition or a promotion was to undermine or eliminate another's opinion or position. Corporate executives, unsure of their own standing, have removed valuable employees from their teams as employees appeared to have talents not held by the executive. Instead of practicing teamwork and encouraging value for contributions, many employees have been taught to divide and conquer within their own departments and, ultimately, within the company. Secretive meetings, positioning, and last-minute layoffs have been common during the past decade.

Win/lose attitudes, common in many large and small corporations, created unhealthy models for competition or for partnering. Less than ethical personal and corporate behavior has had an enormous backlash. Employees laid off or rearranged in new positions many times coincidentally end up working together again. Former managers have become employees of individuals the manager once fired. Competitors purchasing formerly competing companies have forced a merging of teams. It seems that in many cases "what goes around, comes around."

Winning through Intimidation (Ringer, 1993) suggested that if one learns how to intimidate first, then one might be able to avoid the pitfalls of being intimidated. It seems

that in this world of global competition and technology, if one is going to use intimidation tactics to negotiate a partnership, they probably are too weak to be considered a partner for anyone. The best partners come from strength, whether it is personal, financial, spiritual, or corporate success. Finding the strongest partner and then identifying strengths that could be brought to complete them is the best way to formulate a long-lasting strategic alliance. The partners that continually help each other to improve and grow are those that last. In the end, the stronger the partners, the greater enjoyment and financial success for all involved.

Parts of the Contract Never to Forget!

The intent of a contract and the payment arrangements and revenue sharing is usually spelled out from the beginning. Most companies and individuals are anxious to establish criteria for cash assets. Other important assets many contracts tend to forget or to neglect have to do with company identities and branding. Will the new entity have its own identity or will the shared companies have permission to use each other's logos, marketing, and branding? What about the intellectual property? Who owns what and will some ownership be shared? Most technology companies are very familiar with the intellectual property that they own. The problem remains that as new specific partnerships are put into place between two or more companies, the new marketing or technology solutions do not take into account how the solutions will be used by each of the involved parties.

CollegeCapital had cobranded with an international firm to market specific blended products and assumed that the cobranded name of the effort had been checked and trademarked by the dozens of attorneys that had been involved in the formation of the products and services. Several months and millions of marketing dollars later, a competitor company was noticed as using the same name in a competitive situation. After several letters, negotiations, etc., the entire program had to be renamed and redesigned by the partners because the proper legal documents and protections had not been put into place and the new company had created a claim to the branding and names that had been created. A seemingly, slight oversight became a very costly mistake.

Remember: Do not assume that because a company has millions of dollars of revenue, a huge market share, and thousands of employees that all of the appropriate measures have been addressed. Sometimes the most obvious issues are overlooked.

As simple as it sounds, a list of all possible mishaps should accompany all contracts, and at the very least, be items for discussion. There is no way to predict the future with perfect accuracy. This is again where the spirit of the individuals involved and their commitment to solidifying a win/win long-standing partnership becomes the most serious of considerations.

Closing the Deal

It sometimes seems that the negotiation processes continues long past the time of the hoped for contract closing. In several cases, the demand and timing for the alliance partnership may have been so great that the partnership started working before the paperwork was complete. This is another caution in ensuring that your partners will uphold their promises. Whenever possible, don't start working on something without the proper agreements and contracts in place. It usually causes fewer problems to wait a few days and do whatever it takes to complete the contracts.

Creating a Process for Maintenance and Growth of the Strategic Alliance

A simple strategy for managing any relationship is to agree on specific needs for communication as early as possible. Simplifying all processes to be easy to duplicate throughout all of the companies will hasten the success of any project but especially the relationships between individuals within several companies. Look at the communication and process designs as if a franchised system was being designed for the least knowledgeable employee.

The following questions should be answered whenever any consideration to design the implementation of creating a strategic alliance occurs:

What is the ultimate long-term goal to be established using this particular procedure?

What is the short-term result that we should be able to quantify?

Is there a simpler way to achieve this outcome?

What are the quantifiable results that should be expected in one week, one month, 1 year, 5 years, and 10 years?

If all goes well, where will the company be in 20 years?

What reassessment strategies will be used and how often?

What are ways to consider now for growing the relationship for more profits into the future?

CONCLUSION: STRATEGIC ALLIANCES ON A LARGER PLATFORM SUCH AS THE NATION OR THE WORLD

If individuals focused on developing the best of their talents and trusted that they could find others—businesses and persons with whom they could form collaborations to create a unique dynamic product or result—would that not eventually create a better, more peaceful world? Wouldn't companies, countries, or persons be less likely to destroy that which they need for their own gain? And if that need, search, and respect for unique talents were the driving force throughout society, wouldn't we be compelled to nurture, educate, and create opportunities to seek out the uniqueness in all persons, companies, governments, schools, and even religions? Could this constant seeking for completion begin to honor our own seeds of greatness? Wouldn't we be perpetual in offering what we are really great at and admitting to others what we seek or need? Every family, society, and company reflects the thoughts and attitudes of its leaders.

In the Current Events section of *Forbes Magazine* of July 8, 2002, p. 43, Ernesto Zedillo, former President of Mexico, wrote about a new round of trade liberalization talks. He said that during the last talks,

> stressing the strategic value that the substance and timing of this new WTO round—beyond the significant potential economic effects—have for international stability and security.... The industrialized countries, and particularly the U.S., told the world that constructive interdependence (rather than isolationism) and concerted multilateralism (rather than aggressive unilateralism) are essential components of the answer to the challenges that became brutally evident on Sept. 11. It seemed to be the beginning of a coalition, broader than the military one that would fight decisively against the dangerous polarization between the haves and have-nots of the world. This position was reinforced at the U.N. Financing for Development conference in March. (Zedillo, 2002)

Scandals and corporate bankruptcy uncovered in 2002 within several corporations have completely disintegrated investors' faith, as the public found misuse of money, information, and accounting and business practices. Lack of integrity and trust abounded in the disclosures that surrounded the companies' downfalls. Within the e-commerce environment, bad news spreads faster and farther than ever. If financial success is the ultimate goal, doing things right and partnering with those who do things right is all that works. In the advertising campaign developed for AEP, American Electric Power, the banner reads, "Strength + Agility = Performance," a perfect explanation of what the right strategic alliance partnerships in the e-commerce are able to create.

GLOSSARY

Alliance An association to further the common interests of the members.

Acquisition The act of contracting, assuming or acquiring possession of something.

Antitrust Opposing or intended to regulate business monopolies; usually a legal action brought against parties charged with limiting free competition in the marketplace.

Application service provider (ASP) A managed application/software hosting, usually on the Internet.

Barter To trade by exchanging one commodity for another.

Call centers Provide (usually) 24-hour telephone (and e-mail) support services providing a company system backup for customers placing orders; usually providing a complete solution that allows companies to generate a presence on the Web to fulfill orders from all over the world, and staffing customer service personnel.

Charrette A final, intensive effort of a group of experts to finish a project before a deadline (from the word "chariot," as in the speed of wheels).

Commodity A mass-produced, unspecialized, useful product.

Core competence A primary area of expertise; narrowly defined fields or tasks at which a company or business excels; primary areas of specialty.

Customer-centric Focused on or revolves around a customer.

E-commerce The business of buying and selling goods and services on the Internet.

Executive time-share To use executives on an as-needed basis, simultaneous to the use of the same executives by other companies or organizations.

Franchise A license granted to an individual or group to market a company's goods or services in a particular territory.

Golden rule "Whatever you wish that men would do to you, do so to them" (Christian Bible, *Matthew:* 7.12).

Joint venture A partnership or conglomerate formed often to share risk or expertise.

Merger The union of two or more commercial interests or corporations.

Multilateralism Involving or participated in by more than two nations or parties.

Paradigm shift A shift of complete transformation from what was to what is and will be. It is a change in thinking, a revolution, or a sort of metamorphosis. It is usually driven by agents of change.

Partnership A legal relationship between two or more persons contractually associated as joint principals in a business.

Strategic Of great importance and emphasized within an integrated whole or to a planned effect.

Strategic partner Person or company you are closely involved with in some way for a planned effect.

Subcontractor A person or company that does part of a job that another person or company is responsible for.

Tactical partner One that is calculated, clever, deliberate, planned, political, prudent, shrewd, strategic, or skillful in planning or maneuvering under pressure to accomplish a purpose.

Think-tank A consortium of stakeholders and idea people who by coming together in a sequestered, specific time-allotted, environment develop ways to formulate new plans or solve issues.

Unilateralism Done or undertaken by one person, nation, or party, or relating to or affecting one side of a subject.

Vendor One that sells.

World Trade Organization (WTO) A rules-based, member-driven organization—all decisions are made by the member governments, and the rules are the outcome of negotiations among members.

CROSS REFERENCES

See *Application Service Providers (ASPs); Electronic Commerce and Electronic Business; Web Hosting.*

REFERENCES

Asma, G. S. R. (1992). Strategic alliances: Underlying factors, shifts and implications. *ASCI Journal of Management, 21* p. 5-7.

Boeing (1997, October 13). *Boeing, Lockheed Martin join forces to offer U.S. Army streamlined logistics and upgrades for Apache* (news release). Retrieved from http://www.boeing.com/news/releases/1997/news_release_971013o.html

Carnegie, D. (1936). *How to win friends and influence people*. New York: Simon & Schuster.

Cisco Systems (2002a). Other Cisco programs: Introduction. Retrieved from November 3, 2002, from http://www.cisco.com/en/US/partners/pr46/partners_pgm_category_page.html

Cisco Systems (2002b). Strategic alliances. Retrieved April 6, 2003, from http://www.cisco.com/en/US/partners/pr67/part_strat_alliance_category.html

Corporate values: A shift toward collaboration (n.d.). CPFR. Retrieved April 12, 2003, from http://www.cpfr.org/CorporateValues.html

Current process state (n.d.). CPFR. Retrieved April 13, 2003, from http://www.cpfr.org/CurrentProcess.html

Rackham, N., Friedman, L., & Ruff, R. (1996). *Getting partnering right: How market leaders are creating long-term competitive advantage*. New York: McGraw-Hill.

Ringer, R. J. (1993). *Winning through intimidation,* New York Fawcett Crest Books.

Wilson, L. (1994). *Stop selling start partnering*. New York: Wiley.

Webster's New Collegiate Dictionary (1977). Springfield, MA: Merriam.

Zedillo, E. (2002, July 8). Current events: All politics is global, too. *Forbes*, 43.

FURTHER READING

Amor, D. (2002). Internet future strategies: How pervasive computing services will change the world. Englewood Cliffs, NJ: Prentice Hall.

Campbell, A., & Sommers-Luch, K. (1997). *Core competency-based strategy*. London: International Thomson Business Press.

The Golden Rule (n.d.). Retrieved June 10, 2002 from http://www.fragrant.demon.co.uk/golden.html

Greenstein, M., & Vasarhelyi, M. (2002). *Electronic commerce security, risk management, and control*. New York: McGraw-Hill/Irwin.

Gutzman, A. (2002, January 15). *Insights—Ebusiness illuminator: The value of strategic partnerships*. Retrieved from http://ecommerce.internet.com/news/insights/ebiz/print/0,10379_955541,00.html

Hendrick, T. (1997). *Purchasing consortiums: Horizontal alliances among firms buying common goods and services: What? Who? Why? How?* Retrieved from http://www.capsresearch.org/ReportHTMs/consortiums.html

Inkpen, A. C., & Ross, J. (1996, October 1). Why do some strategic alliances persist beyond their useful life? *California Management Review. 39*(1), 123–140.

ITEX in Florida. (n.d.). *Frequently asked questions*. Retrieved May 16, 2002, from http://www.itexflorida.com/faq.htm

Lendrum, T. (1998). *The strategic partnering handbook*. Sydney: McGaw-Hill.

Nelson, C. R., & Saltiel, D. H. (2002, January). *Managing proliferation issues with Iran* (policy paper). Washington, DC: Atlantic Council of the United States.

Pacific Coast Reproductive Society. (2001). *Strategic partnerships PCRS*. Retrieved June 26, 2002 from http://www.pcrsonline.org/strategic_partners.htm

Rehder, R. R. (1989, Winter). Japanese transplants: In search of balanced and broader perspective. *Columbia Journal of World Business*, 17–28.

Rigsby, E. (2000). *Developing Strategic Alliances*. Los Altos, CA: Crisp.

United States Department of Justice (2002, May 26). *Antitrust enforcement and the consumer*. Retrieved from (July 23, 2002) http://www.usdoj.gov/atr/public/div_stats/1638.htm

Voluntary Interindustry Commerce Standards (VICS) (2002, June). Introduction. In *Global commerce initiative: Recommended guidelines: Collaborative planning, forecasting and replenishment (CPFR), Version 2.0*. Retrieved July 20, 2002, from http://www.cpfr.org

Wood, R. C., & Hamel, G. (2002, November 1). The World Bank's innovation market. *Harvard Business Review*.

Structured Query Language (SQL)

Erick D. Slazinski, *Purdue University*

INTRODUCTION

Applications developed today, including Internet applications, data-mining software, and other general applications, will most likely have database components on the back end. Example applications vary from the online CD database located at CDDB.com to a recipe database, like the one located at FOODTV.COM. The databases used may range from powerful enterprise-scale versions of Oracle, DB2, and Sybase to more personal desktop version like Microsoft Access. This chapter will focus primarily on standard SQL, not any specific implementation. Regardless of the database platform, one accesses the data utilizing the provided query language. Because most popular databases use the relational model, the query language provided is a derivative of SQL. Currently most database systems provide users with languages such as SQL to allow them to retrieve or update stored information (Chan, Tan, & Wei, 1999).

There are several mechanisms for a user to interact with a database. Most database vendors provide users with line-mode interfaces where a user can enter supported statements and receive results. Other, graphical interfaces may be available from a database vendor or a third-party application. These graphical interfaces provide what is known as query-by-example (QBE) facilities. QBE was developed by IBM in the 1970s to help users in their retrieval of data (Connolly & Begg, 2002, p. 197). The QBE facilities provided by database vendors either allow users to graphically depict queries or provide templates for users to fill out. These facilities then generate SQL, which often can be viewed and modified and which then is sent to the database engine for execution.

MATHEMATICAL BEGINNINGS

Although database technology has been around for quite some time and in various forms, the relational model put forth in 1970 by E. F. Codd has enjoyed the greatest success. The relational model is firmly grounded in relational algebra, which has roots in basic set theory. This actually makes the basics of SQL accessible to most people.

Basic Set Theory

In basic set theory, a universe of domain is specified. In the database arena this translates to all of the data included in a database. In our universe, items may be categorized (perhaps multiple times) into set(s) and then relationships between these sets may be explored, using inclusion and exclusion operations. Likewise, in the relational model, data are often retrieved as result sets. These result sets may then be compared either inclusively or exclusively.

Relational Algebra

Relational algebra "is a collection of operations that are used to manipulate entire relations. . . . The result of each operation is a new relation, which can be further manipulated by the relational algebra operations" (Elmasri & Navathe, 1989, p. 148). This is consistent with the mathematical definition of set operators. Relational operators are often separated into two groups. The first "group includes set operations from mathematical set theory . . . . The others consist of operations developed specifically for relational databases . . . ." (Elmasri & Navathe, 1989, pp. 148–149).

The set of relational algebra operators includes the SELECT (σ), PROJECT (π), UNION ($\cup$), DIFFERENCE ($-$), and CROSS-PRODUCT ($\times$) operators. "It has been shown that the set of relational algebra operations $\{\sigma, \pi, \cup, -, \times\}$ is a *complete* set; that is, any of the other relational algebra operations can be expressed as a *sequence of operations from this set"* (Elmasri & Navathe, 1989, p. 159). An example of a complex relational operator (one made from this minimal set) is the intersection operator ($\cap$), which is defined as

$$R \cap S \equiv (R \cup S) - ((R - S) \cup (S - R))$$

(Elmasri & Navathe, 1989, p.159).

"The select operation is used to select a subset of tuples (tuples represent rows in a table) in a relation. These tuples must satisfy a *selection condition*" (Elmasri & Navathe, 1989, p. 149). The general form for the SELECT operation is $\sigma_{<\text{selection\_condition}>}(<\text{relation name}>)$, where the SELECTION_CONDITION is a boolean expression made up of clauses of the form <attribute name> <comparison op> <constant value> or <attribute name> <comparison op> <attribute name>, where ATTRIBUTE NAME is the name of an attribute of the <relation name> (Elmasri & Navathe, 1989, p. 149). For ordered domains, such as integer or date domains, <comparison op> is from the set $\{=, <, >, <=, >=, \neq\}$. For unordered domains, such as color = {red, blue, green}, the only valid <comparison op> is from the set $\{=, \neq\}$. Boolean operators AND, OR, and NOT are used to connect the selection condition clauses together. An example of a table of tuples is as follows:

name	sex	address
Smith	M	Lexington Park
Taylor	M	Detroit
Burr	M	Lusby
Malcolm	M	Akron
Tefft	M	Ridge
Carpenter	M	Lexington Park
Lucas	M	Hollywood

An example of the selection operation is as follows:

```
Give me all candidates who are male.
[select] σ(sex="M") (CANDIDATE)
```

The PROJECT operation is used to select a subset of attributes in a relation. The general form for the projection operator is $\pi_{<\text{attribute\_list}>}(<\text{relation name}>)$. Only attributes that are part of the relation are valid in the attribute list. If duplicate tuples appear in the result relation, all but one instance of the tuple will be removed, enforcing the mathematical set definition, which allows no duplicate items.

An example of the projection operation:

```
Where do the candidates live?
[project]πAddress (CANDIDATE)
```

address
Lexington Park
Washington DC
Detroit
Hollywood
Lusby
Pittsburgh
Cherry Valley
Akron
Ridge

As stated previously, all relational operators may operate on the results of previous relational operators.

Examples involving both select and project operators:

```
Give me all of the names and addresses of
  the candidates that live in Hollywood.
πName, Address (σ(address = "Hollywood") (CANDIDATE))
```

name	address
Rupert	Hollywood
Day	Hollywood
Lucas	Hollywood

The above relational operators are unary, operating on only one relation at a time. The rest of the relational operators are binary, operating on two relations at a time. The UNION and DIFFERENCE operators require that the relations also be union-compatible. The UNION operation returns all tuples (rows) that are either in R or S, or in both R and S. Enforcement of the set definition eliminates duplicate tuples. The difference operation returns all tuples that are in R but not S. The CARTESIAN PRODUCT operation returns all possible tuple combinations between R and S. The resulting relation has all attributes from both relations and the new tuples are of the form $(\text{tuple}_{1_R}, \text{tuple}_{1_S})$, $(\text{tuple}_{1_R}, \text{tuple}_{2_S})$, $(\text{tuple}_{2_R}, \text{tuple}_{1_S})$, $(\text{tuple}_{2_R}, \text{tuple}_{2_S})$, . . . $(\text{tuple}_{m_R}, \text{tuple}_{n_S})$.

The JOIN operator (ξ) is used to combine related tuples from two relations into single tuples (Elmasri & Navathe, 1989, p. 157). The JOIN returns all possible tuple combinations from R and S where the combination satisfies the JOIN CONDITION. An example JOIN is $R\ \xi_{<\text{join condition}>}\ S$. JOIN CONDITIONS are of the form $A_R \theta A_S$, where A_R and

A_S are domain-compatible attributes from relations R and S, respectively, and θ is one of the following comparison operators: $\{=, <, \leq, >, \geq, \neq\}$. JOIN CONDITIONS may be connected with the Boolean AND operator. For example, to retrieve all of the data about a candidate the statement CANDIDATE $\xi_{(\text{candidate.name} = \text{candidate-body.name})}$ CANDIDATE-BODY would yield the following relation:

name	sex	address	name	height (in)	age
Smith	M	Lexington Park	Smith	75	50
Jones	F	Washington DC	Jones	77	26
Taylor	M	Detroit	Taylor	65	36
Rupert	F	Hollywood	Rupert	62	30
Burr	M	Lusby	Burr	71	25
Zimmerman	F	Pittsburgh	Zimmerman	61	38
Roberts	F	Cherry Valley	Roberts	68	40
Malcolm	M	Akron	Malcolm	72	20
Tefft	M	Ridge	Tefft	70	29
Carpenter	M	Lexington Park	Carpenter	74	30
Day	F	Hollywood	Day	64	25
Lucas	M	Hollywood	Lucas	70	19

Null Values and Trivalue Logic

One concept that was added to basic set theory was the concept of null value. An attribute that has a value of null is said to be an attribute that is not applicable, unknown, or undefined for the data row. One must be careful to properly account for the possibility of null values appearing in a database.

Null values, when used in numeric computations, act like infinity; anything we do in conjunction with a null value ends up as null. For example, if we add 3 + null, the result is null. When used in character operations (such as concatenation) the null typically acts like a blank.

The two places where one must handle the null are in the *select* statement and the *where* clause. In the select statement, we must decide what to display to the final recipient of the SQL statement that we are inputting instead of the null, which typically displays as a blank character. In the *where* clause, the results are more profound. In trying to evaluate a *where* clause where portions of the clause may null out (return a value of null), the desired result set may not be returned. This is due to the fact that SQL uses a trivalue logic set as shown in the table below:

p	q	p AND q	p OR q	NOT p
True	True	True	True	False
True	False	False	True	False
True	Null	Null	True	False
False	True	False	True	True
False	False	False	False	True
False	Null	False	Null	True
Null	True	Null	True	Null
Null	False	False	Null	Null
Null	Null	Null	Null	Null

So if one portion of a *where* clause nulls out, depending on the operation, the entire *where* clause may null out, leaving the empty set as our result set. For the "and" operation to return a value of "True" both predicates ("p" and "q" above) must evaluate to "True." Because the value of null is unknown (it could be "True"), the evaluation cannot be complete and therefore returns a null value. For the "or" operation, only one of the predicates must evaluate to "True." The not operation simply changes the predicate's value: a "True" value become a "False" value, and so on.

IBM'S SEQUEL LANGUAGE

Even though Codd proposed relational databases in 1970, it was not until 1974 that Chanberlin and Boyce published an article proposing the form of a structured query language, named SEQUEL (Hursch & Hursch, 1988, pp. 1–2). This name is often used (incorrectly) today to reference the present-day SQL language.

STRUCTURED QUERY LANGUAGE

All of the basic concepts of relational algebra are present in today's relational database management systems (RDBMs). However, the concept of a domain is only implemented for primitive data types, such as integer, float, date, versus any data type, as defined in the model (Date, 1994, p. 66). SQL is the American National Standards Institute (ANSI) standardized version of IBM's SEQUEL "data sublanguage," for use in a relational database (Hursch & Hursch, 1988, p. 1). SQL (pronounced "ess, queue, ell") is more than a language for retrieving information out of tables; it also "includes features for defining the structure of the data, for modifying the data in the database, and specifying security constraints. Each feature has its own set of statements that are expressed in, respectively, Data Definition Language (DDL), Data Manipulation Language (DML), and Data Control Language (DCL)" (Hursch & Hursch, 1988, p. 1). DML implements the relational operators and will be addressed later.

Sets of data, relations, are stored in a RDBMS in the form of tables. Each table consists of rows and columns. Each row of data is a specific and unique instance (entity) of a member of the table (i.e., a tuple). Each column is said to be an attribute of the table. To create the database objects required, SQL's DDL provides the following data definition statements (DDS): CREATE TABLE, CREATE INDEX, ALTER TABLE, DROP TABLE, DROP VIEW, and DROP INDEX. The table names created with these statements are the identifiers used by the DML to resolve SQL statements.

SQL's DML supports the following operations on tables: INSERT, UPDATE, DELETE, SELECT, and CREATE VIEW. INSERT, UPDATE, and DELETE operations are data maintenance functions and will not be addressed. The SQL SELECT is the data retrieval mechanism. The CREATE VIEW statement is an alternative way of looking at the data in one or more tables called a derived table and is based upon the SELECT statement.

SQL's DCL is used to control authorization for data access and auditing database use. DCL statements include the GRANT and REVOKE statements. Database administrators and developers primarily use these operations.

ANSI Standardization

SQL is in its third revision as an ANSI standard. The first revision was released in 1986, the second version (SQL2 or SQL-92) was released in 1992, and the latest release (SQL3 or SQL-99) occurred in 1999. Each release added major functionality to the standard and, in theory, increasing compatibility for those willing to program in "pure" SQL (Eisenberg & Melton, 2002).

> IBM developed SQL in the 1970s for use in System R. It is the de facto standard as well as being an ISO and ANSI standard. It is often embedded in general-purpose programming languages.
>
> The first SQL standard, in 1986, provided basic language constructs for defining and manipulating tables of data; a [minor] revision in 1989 added language extensions for referential integrity and generalized integrity constraints. Another revision in 1992 provided facilities for schema manipulation and data administration, as well as substantial enhancements for data definition and data manipulation.
>
> Development is currently underway to enhance SQL into a computationally complete language for the definition and management of persistent, complex objects. This includes: generalization and specialization hierarchies, multiple inheritance, user defined data types, triggers and assertions, support for knowledge based systems, recursive query expressions, and additional data administration tools. It also includes the specification of abstract data types (ADTs), object identifiers, methods, inheritance, polymorphism, encapsulation, and all of the other facilities normally associated with object data management. (FOLDOC, 2002)

DDL

SQL's DDL provides the following data definition statements (DDS): CREATE TABLE, CREATE INDEX, CREATE VIEW, ALTER TABLE, DROP TABLE, DROP VIEW, and DROP INDEX. The create statements create the named structures within the database (tables, indexes, etc). The drop statements remove the corresponding structures. The create view statement will be handled in a later section.

DML

SQL's DML supports the following operations on tables: INSERT, UPDATE, DELETE, SELECT, and CREATE VIEW. The SQL SELECT is the data retrieval mechanism. The operations are self-explaining; the insert operation adds data into a named table, the update operation modifies data in a named table, and the delete operation removes data from the named table. All of these operations are permanent modifications to the database if and only if they are followed by commit statements—refer to the section on transaction control.

The Select Statement

The select statement is one of the most flexible and often used in SQL. The simplest form of the SELECT statement, "select * from <table name>," returns all of the rows from the indicated table. By specifying a *where* clause, the resultant set is restricted by a set of conditions. The *group by* clause allows data grouping (ordered by the attributes listed). The *having* clause can further restrict the results returned by the *where* clause. The *order by* clause allows a final ordering to occur before results are returned to the requester. The SQL SELECT statement has no restrictions on the number of tables that may be accessed to fulfill the user's request. Besides being the query construct, the select statement can be used in conjunction with every DML statement. The syntax of the select statement is

```
select      * | <attribute list>
from        <table list>
{where      <condition> }
{group by   <grouping attributes> }
{having     <group condition> }
{order by   <attribute list> }
```

SINGLE TABLE ACCESS AND THE RESULT SET

Because SQL was designed using set theory, one important carry-over was the result set. The result of any query is a set, named the result set. This makes for a very powerful and flexible design. Because we have a set, we can then apply another query to this result set (in the case of refining as we go).

The most basic of queries returns the contents of a specified table:

```
select * from TICKET_PRICE;
```

The result set is the entire ticket_price table:

TICKET_TYPE	PRICE
KIDS	4.50
SENIORS	5.00
ADULT MAT	5.00
ADULTS	7.00

The "select *" statement is often used to inspect the data that are stored in a table (that is not too large) and to verify that DML operations were successful (especially inserts and deletes).

Reducing the Size of the Result Set

Often the "select *" statement returns too many rows for useful analysis. To trim down the size of the result set, we can perform some combination of column and row reduction. To reduce the number of columns, we simply replace the "*" with the column names that we want to investigate. To reduce the number of rows, we must include a *where* clause. The *where* clause provides a list of criteria that can be joined using trivalue Boolean logic.

The criteria specified can be very specific, such as returning those rows that have this exact value for an attribute (e.g., returning all movies with a rating of "PG"). The criteria can be less specific (e.g., returning all movies that begin with an "A"). Examples follow:

To return all movie titles with a PG rating (exact match):

```
select MOVIE_TITLE
from MOVIE
where MOVIE_RATING = "PG";
```

The result set would be

MOVIE_TITLE
THE BEST MAN IN GRASS CREEK
CROCODILE DUNDEE IN L.A.
SPY KIDS

To return all movie titles and rating which begin with the letter "A" (inexact match):

```
select MOVIE_TITLE, MOVIE_RATING
from MOVIE
where MOVIE_TITLE like 'A%';
```

The result set would be

MOVIE_TITLE	MOVIE_RATING
A KNIGHT'S TALE	PG-13
AMORES PERROS	R

Note: from this point forward, unless difficult to determine, the results sets will be left as an exercise to the reader.

To return all movie titles with a PG or PG-13 rating (exact match), we can use either exact or inexact matching techniques. If a new movie rating was created, say PG-15, and we only wanted PG or PG-13 movies, the inexact technique would no longer work. The following queries will return the same result set:

```
select MOVIE_TITLE
from MOVIE
where MOVIE_RATING in (
       "PG," "PG-13")
```

```
select MOVIE_TITLE
from MOVIE
where MOVIE_RATING
like 'PG%';
```

```
select MOVIE_TITLE
from MOVIE
where MOVIE_RATING ='PG'or
      MOVIE_RATING ='PG-13';
```

Other queries using string searching techniques are also possible.

If we want to get a range of values, we can use the *between* keyword. For example, we want all movie tickets priced between $5.00–$7.00. The query would look like this:

```
select *
from TICKET_PRICE
where PRICE between 5.00 and 7.00;
```

The following comparison operators are specified for use in a *where* clause: <, >, =, ! =, <>. IN, BETWEEN, and LIKE are used on string data only.

Computations

Perhaps we want to enlarge the result set or perform a computation, such as computing the total for a shopping cart application or performing what-if scenarios relating to the price of a movie ticket. We can perform operations on the attributes listed in the select statement. Note: we cannot perform these operations when we use the "select *" notation. We can either perform a static computation (e.g., what would our pricing structure look like if we raised ticket prices by 15%?) or a dynamic computation (e.g., a calculation involving two or more attributes). An example query would be:

```
select TICKET_TYPE, PRICE, PRICE * 1.15
from TICKET_PRICE;
```

Arranging the Result Set

SQL provides us with a mechanism for sorting the rows of a result set. This mechanism is the *order by* clause. When using the *order by* clause, the query developer specifies which attribute(s) to sort the data on and the direction asc (ascending) or desc (descending). If more than one attribute is specified, then the outer attribute specifies the major ordering and the next attribute(s) indicates the ordering within the primary ordering.

Note: the *order by* clause is the last clause of a select statement. For example, the following query using the *order by* clause generates the result set that follows:

```
select MOVIE_RATING, MOVIE_TITLE
from MOVIE where MOVIE_TITLE like 'A%'
order by MOVIE_RATING desc, MOVIE_TITLE asc;
```

MOVIE_RATING	MOVIE_TITLE
R	ALONG CAME A SPIDER
R	AMORES PERROS
PG-13	A KNIGHT'S TALE

Segregating the Result Set

Often we want to work on a partitioned result set (i.e., for each movie rating in our result set, we want to count the number of movies that have the rating). Or we want to get the most (or least) expensive book in our inventory. To support these types of queries, SQL has defined a number of aggregate functions. Example aggregate functions include COUNT(), MIN(), MAX(), SUM(), and AVG(). All functions take in column expressions (either column names or column computations). The count function can also take in a "*." In order, these functions perform the following computations—counting the number of rows, determining the minimum value, determining the maximum value, totaling the value, and determining the average value. In general, these functions are most useful when we partition the result set (e.g., total

value of inventory by product line) as opposed to running them on an entire result set (such as an entire table's contents).

Examples

How many movies of each rating are there?

```
select count(*), MOVIE_RATING
from MOVIE
group by MOVIE_RATING;
```

Results:

COUNT(*)	MOVIE_RATING
3	PG
6	PG-13
11	R

The various functions can be used without the *group by* clause; however, the value is computed for the entire table. The COUNT(*) function (without the *where* clause) is useful to determine the number of rows that a table contains.

Reducing the Segregated Data

Now that our data are partitioned and we have some calculated information about them, we may want to reduce the number of rows that are output. The proper way to accomplish this is to use the *having* clause. If we continue with the above example, and now only want to view the data associated with the PG and PG-13 grouped values, we could add a *having* clause like this:

```
select count(*), MOVIE_RATING
from MOVIE
group by MOVIE_RATING
having MOVIE_RATING like 'PG%';
```

Note: the *having* clause must be used in conjunction with a *group by* clause, though a *group by* clause can exist without a *having* clause.

Handling Null Values

Remember that null is a special value that can be assigned to an attribute. Most functions, like COUNT(), will ignore a column with a null value. Other calculations, like mathematical operations, treat null as infinity, so that when we add a number to a null value, the result is also a null—that is the result is unknown. There may be times when we want to search out these unknown or not applicable values to fulfill some business query. To search for an attribute with a null value we cannot use the " = " comparison operator; we must use the keyword *is*, yielding an expression of the form "MOVIE_RATING is null." Of course, the converse to "is null" exists: it is "is not null." Many RDBMSs also provide replacement mechanisms for generating reports, so that the null value does not interfere with calculations and the like. Refer to an RDBMS's user guide for the exact syntax.

Joins

So far the various options have been demonstrated against a single table, but if we are using a relational database, then the relationships between tables are equally important. To retrieve result sets with data from multiple tables, we need to use a join. There are several types of joins defined: equi, natural, Cartesian, inner, outer, and self. The equi and natural joins are the most common; the inner, outer, self, and Cartesian are special cases. In all cases, all tables that participate in a join must be listed in the *from* clause.

Equi-joins

An equijoin combines the data from two or more tables based on a match defined in the *where* clause. For example, if we wanted to count the number of tickets sold for a particular movie, we would enter:

```
select count(*)
from TICKETS_SOLD, MOVIE
where TICKETS_SOLD.MOVIE_ID = MOVIE.
  MOVIE_ID and
       MOVIE_TITLE = 'DRIVEN';
```

Natural

A natural join is a specific instance of an equijoin. The join columns must have the same name. The fields must be the same length and data type. When the result set is generated, only one instance of the join column is included.

Cartesian

The product of a Cartesian join is all possible combinations of data combined in one result set. If we have a table A with three rows and a table B with two rows, the Cartesian product result will have six rows:

A
A1
A2
A3

B
B1
B2

select * from A,B
A1 B1
A1 B2
A2 B1
A2 B2
A3 B1
A3 B2

Union

The union operation combines the result sets of two or more queries together into a new result set. This corresponds directly with the union operation in set theory. The only restriction is that the result sets being joined must be union-compatible. That is, the numbers of columns in the result sets to be joined must be the same. Additionally, every corresponding column must be of the same datatype (that is, column 1 of result set 1must be the same datatype as all of the other column 1s from all of the other result sets being so joined). It is up to the developer to ensure that the meaning of the columns is consistent between result sets. Unions are useful when combining related data

that are stored in different tables—such as salaried employee wages and hourly employee wages. It makes no sense to have number of tickets and ticket price joined via a union, but SQL will probably allow it because they are both numeric columns. The syntax is:

```
select statement 1
UNION
select statement 2
```

Minus

The minus operation, also directly ported from set theory, removes matching rows from multiple result sets. This is useful in determine what is different between these results—or "what didn't sell while it was under a promotion." The syntax is:

```
select statement 1
MINUS
select statement 2
```

DDL

Create Table

The *create table* statement allocates storage for data to be stored. This storage has a structure (columns with datatypes) and optional constraints defined. Once the table is created, the table is ready to receive data via the INSERT statement (see below). The syntax is straightforward:

```
create table <TABLE_NAME> (
     <column_element> | <table_constraint> )
```

For example, to create the movie table located in Appendix A, the following syntax was used:

```
CREATE TABLE MOVIE (
       MOVIE_ID         NUMBER        NOT NULL,
       MOVIE_TITLE      VARCHAR2(50)  NOT NULL,
       INVENTORY_ID     NUMBER        NOT NULL,
       MOVIE_RATING     CHAR(5)       NOT NULL,
       PASS_ALLOWED     CHAR(1)       NOT NULL);
```

Keys

When a database is modeled using entity–relationship techniques, identifying attribute(s) are identified for all fundamental entities (tables), along with those attributes that relate two or more tables. The identifying attribute(s) uniquely distinguish one row of data from all others in the table. In SQL these identifying attribute(s) are identified to the database engine as a primary key. Many database engines use this information to enforce the uniqueness property that is inherent in the definition. Attributes that tie relations together are known as foreign keys. Keys can be identified either when a table is created (with the create table command) or after (via the alter table command).

Create Index

The create index statement allows the database engine to create an index structure. An index structure provides a mapping to data in a table based upon the values stored in one or more columns. Indexes are used by the database engine to improve a query's performance by evaluating the *where* clause components and determining if an index is available for use. If an index is not available for use during query processing, each and every row of the table will have to be evaluated against the query's criteria for a match. The syntax is not part of the SQL language specification (Groff & Weinberg, 1999, p. 387) though most DBMSs have a version of the create index statement.

For example, to create an index on the MOVIE_TITLE of the movie table located in the Appendix, the following syntax was used:

```
create index MOVIE_NAME_IDX on MOVIE
  (MOVIE_TITLE);
```

Create View

"According to the SQL-92 standard, views are virtual tables that act as if they are materialized when their name appears" (Celko, 1999, p.55). The term *virtual* is used because the only permanent storage that a view uses is the data dictionary (DD) entries that the RDBMS defines. When the view is accessed (by name) in a SQL *from* clause, the view is materialized. This materialization is simply the naming, and the storing in the RDBMS's temporary space, of the result set of the view's select statement. When the RDBMS evaluates the statement that used the view's name in its *from* clause, the named result set is then referenced in the same fashion as a table object. Once the statement is complete the view is released from the temporary space. This guarantees read consistency of the view. A more permanent form of materialized views is now being offered by RDBMS vendors as a form of replication, but is not germane to this discussion.

The syntax for creating a database view is below:

```
create view <name> [<column list>] as
         <select statement>
```

The power of a view is that the select statement (in the view definition) can be almost *any* select statement, no matter how simple or complex—various vendors may disallow certain functions being used in the select statement portion of a view's definition.

Views have proven to be very useful in implementing security (restricting what rows and/or columns may be seen by a user); storing queries (especially queries that are complex or have specialized formatting, such as a report); and reducing overall query complexity (Slazinski, 2001).

Constraints

Constraints are simple validation fragments of code. They are (from our perspective) either the first or last line of defense against bad data. For example, we could validate that a gender code of "M" or "F" was entered into an employee record. If any other value is entered, the data will not be allowed into the database.

Domains

A domain is a user-created data type that has a constrained value set. For our movie example we could create a movie rating domain that was a 5-character data type that only allowed the following ratings: {G, PG, PG-13, R, NC-17, X, NR}. Once the domain was created, it could be used in any location where a predefined data type could be used, such as table definitions. Not all database vendors support the creation and use of domains; check the user guide for information and syntax specifications.

DCL

SQL's DCL is used to control authorization for data access and auditing database use. As with any application with Internet access, security is a prime concern. Today's RDBMSs provide a host of security mechanisms that the database administrators can use to prevent unauthorized access to the data stored within.

Granting Access to Data

By default, in an enterprise-edition RDBMS, the only user who has access to an object (table, index, etc.) is the user who created that object. This can be problematic when we have to generate a report based on information that is stored in tables that are owned by other users. Fortunately, the solution is the *grant* statement. The syntax is

```
grant <access> on <table> to <user>;
```

One drawback to the grant statement is that it requires a grant statement for every user–object–access level combination—which can be a challenging task in a large enterprise where data security is critical. Accounting departments usually want a high degree of accountability from those individuals who have access to their data. Depending on a database's implementation, there may be mechanisms in place to help ease the burden of security management.

DML

Inserting Data

Syntax:

```
insert into <table> (column {[, column]}))
             values (<value> | <expression>)
```

For A Single Row

For example, inserting a new row of data into the ticket_price table would look like this:

```
insert into TICKET_PRICE values (
                     'MIDNIGHT SHOW', 5.50);
```

For Multiple Rows

If we want to insert multiple rows of data we simply have multiple insert statements; however, if we have data stored in another table that we wish to place in our table (perhaps copying data from a production database into a developer's personal database), we can use an expression.

For instance, if we had ticket price information in a table named master_ticket_price, and we wanted to copy the data into our local ticket_price table, we could issue the following command:

```
insert into TICKET_PRICE
    select * from MASTER_TICKET_PRICE;
```

Now this is assuming that the structure of THE MASTER_TICKET_PRICE table and the TICKET_PRICE table are the same (i.e., the same columns). If we only want a partial set of values from the MASTER_TICKET_PRICE table, we can add a *where* clause to our select statement. Likewise, we can reduce the number of columns by selecting only those columns from the MASTER_TICKET_PRICE table that match our TICKET_PRICE table.

Modifying Data

Syntax:

```
update <table>
set <column> = <value> | <expression>
{where_clause};
```

For a Single Row

To update only one row in a table, we must specify the *where* clause that would return only that row. For example, if we wanted to raise the price of a KIDS movie ticket to $5.00, we would issue the following:

```
update TICKET_PRICE
set PRICE = 5.00
where TICKET_TYPE = 'KIDS';
```

We could perform a calculation or retrieve a value from another database table instead of setting the price equal to a constant (5.00).

For Multiple Rows

For modifying multiple rows there are two options—if we want to modify all of the rows in a given table (say to increase the price of movie tickets by 10% or to change the case of a column), we just leave off the *where* clause as such:

```
update TICKET_PRICE
set PRICE = PRICE * 1.10;
```

If we want to modify selected rows, then we must specify a *where* clause that will return only the desired rows. For example, if we want to raise just the adult prices by 10%, we issue the following:

```
update TICKET_PRICE
set PRICE = PRICE * 1.10
where TICKET_TYPE like 'ADULT%';
```

Removing Data

Syntax:

```
delete <table>
{where_clause};
```

For a Single Row

To delete only one row in a table we must construct a *where* clause that returns only that row. For example, if we wanted to remove the SENIORS movie ticket type, we would issue the following:

```
delete TICKET_PRICE
where TICKET_TYPE = 'SENIORS';
```

For Multiple Rows

For deleting multiple rows there are two options—if we want to delete all of the rows in a given table, we just leave off the *where* clause as such:

```
delete TICKET_PRICE;
```

If we want to delete a set of specific rows from a table, then we must specify a *where* clause that returns only those rows. For example, if we want to delete only the movies that allow passes, we issue the following:

```
delete MOVIE
where PASS_ALLOWED = 'Y';
```

TRANSACTION CONTROL

A database transaction is a unit of work performed. This could be the checkout portion of a shopping cart application—when the user finally hits the purchase button. It could be a data entry person entering time card data or performing an inventory to verify that stock levels match what is on the shelves. In database terms a transaction is a series of DML statements that are logically grouped together. Because there are humans involved, SQL provides us with the ability to recover from mistakes (to a certain degree). We can start a transaction by issuing a "begin transaction" statement. This marks the beginning of our work session. We can then proceed with modifying the data in any way that is required. Once we are sure of our changes—verified via SQL queries, no doubt—we can issue a "commit." The commit statement tells the database engine that we are done with this unit of work and the changes should be made a permanent part of the database. If we are not satisfied with out changes or we made a mistake we can issue a "rollback" statement, which undoes all of the work performed since the "begin transaction" was issued. Note that SQL does not support multiple levels of undo, like many computer applications. Once a commit is issued, we must manually undo any portion of our transaction that we are not satisfied with. The classis example that is often used to illustrate mutual, dependent, or two-phase commits is the use of an ATM machine. When someone uses the ATM machine he or she expects to receive the money requested, and the account will be altered to reflect this withdrawal. Likewise, the bank is willing to give the requesting person his or her money (if sufficient funds are available) and the account is properly adjusted. If both of these actions (dispensing funds and altering account) are successful, then all is well. If either action fails, then all transactions must be rolled back.

MULTIUSER ENVIRONMENTS

One of the reasons for the success of the RDBMS is its ability to handle multiple, simultaneous users. This is a critical feature for many Web sites. Amazon.com would not be around too long if only one person at a time could look up a book's information or check out. However, in order to handle multiple, simultaneous users we must have some locking mechanism in place in order for users not to overwrite each other's information. Likewise, we do not want to delay a user's ability to browse through our dataset while it is being worked on—a common problem for any system (Web site or otherwise) that is in a 24×7 support mode. Various operations cause various types or levels of locks to be placed on the data. For example, in modifying the data, an exclusive lock is entered. This forbids any other user to modify the same data until the first user has completed his or her transaction. A commit or rollback statement will either permanently record or undo the change. This is the why developers should commit often and as early as possible. Too many locks without committing or rolling back the data have profound performance implications. Different RDBMSs lock at different levels—some lock only the data that are modified, others lock certain chunks of data (the affected rows and other rows that are contiguously stored with them). Techniques for dealing with various locks are beyond this chapter.

Concurrency Issues

Concurrency is a big issue in multiuser systems. Every user wants the most current data, yet just because we are entering data into a system does not mean that we entered them correctly—this is why transaction control was invented. All enterprise RDBMSs have a form of locking that permits the querying of data that are in the process of being modified. However, because the data being modified are not yet committed to the database, the query will only be allowed to see the committed data. The impact is this—we can run a query, study the result, and 5 min later, rerun the query and get different results. The only alternative would be to have all queries wait until all of the pieces of information that they are requesting were not being modified by any other users. Another impact, where a user may have to wait until another user has completed a transaction, is when a user wants to modify a piece of information that is in the process of being modified by another user. In this case only one user can modify a piece of information (set of rows in a table) at a time.

Deadlock

Deadlock is a condition where two users are waiting for each other to finish. Because they are waiting on each

other, they can never finish their processing. An example of deadlock would be as follows: User 1 is updating all of the rows in the MOVIE table and at the same time User 2 is updating all of the rows in the TICKET_TYPE table. Now, User 1 decides to update all of the rows in the TICKET_TYPE table. The database will make the user wait (because User 2 currently has the TICKET_TYPE data locked—the transaction has not been committed). Meanwhile, User 2 decides to modify all of the rows in the MOVIE table. Again the database will make User 2 wait (because User 1 has not committed his or her data). At this point deadlock has occurred—both users are in a wait state, which can never be resolved. Many commercial RDBMSs test for deadlock and can perform some corrective measures once it is detected. Corrective measures can range from aborting the SQL statement that caused the deadlock to terminating the session with the least amount of process time that was involved with the deadlock—refer to the RDBMS's user guide for details on a system.

ENHANCED VERSIONS OF SQL

Even though there have been three versions of the SQL language published, there is no vendor that adheres 100% to the standard. All vendors add features to their database engines in an attempt to entice consumers. This is not always a bad feature. This is part of the reason that there have been three standards to date: the database vendors are continuously improving their products and exploring new areas that can be supported by database technologies. Two of the most popular extensions are procedural extensions and specialized derivatives of SQL.

Procedural Extensions to SQL

Although SQL is great at manipulating sets of data, it does have some shortcomings. A few immediately come to mind. It is usually very difficult with standard SQL to produce a top *n* listing, e.g., the top 10 songs on the *Billboard* charts. It is impossible to process a result set, one row at a time. Complex business rules governing data validation cannot be handled by normal constraint processing. Last, performance of SQL queries (from the entire parse, execute, send results perspective) can be very poor when the number of clients is large, such as in a Web environment. Note that a performance gain can be had if every query was stored as a View (see above). The first three of these items are typically handled in stored modules and triggers. The last item can be handled either by using database views or by making use of the database engine's query caching scheme (if it exists)—this is beyond the scope of this chapter.

Stored Modules

A stored module (typically a procedure or function) is a piece of procedurally enhanced SQL code that has been stored in the database for quick retrieval. The benefits of having the code stored are numerous. By storing the code, an end user does not have to type the code in each and every time that it needs to be called upon. This supports the concepts of code reuse and modularity. Performance is increased, because the code has been verified to be correct and security policies can be enforced. For example, if we want to limit access to a given table, we can create an insert stored procedure and grant end users rights to execute the procedure. We can then remove all rights to the table from all users (except the table's owner) and now, if our database is broken into by any username other than the table's owner, the only operation that can be performed is the insert stored procedure.

Triggers

A trigger is a piece of procedurally enhanced SQL code that has been stored in a database, which is automatically called in response to an event that occurs in the database. Triggers are typically considered unavoidable—i.e., if the event for which a trigger has been defined occurs, the trigger fires (is executed). Some RDBMSs do allow bulk load programs to disable the triggering mechanism—refer to the RDBMS's user guide for details.

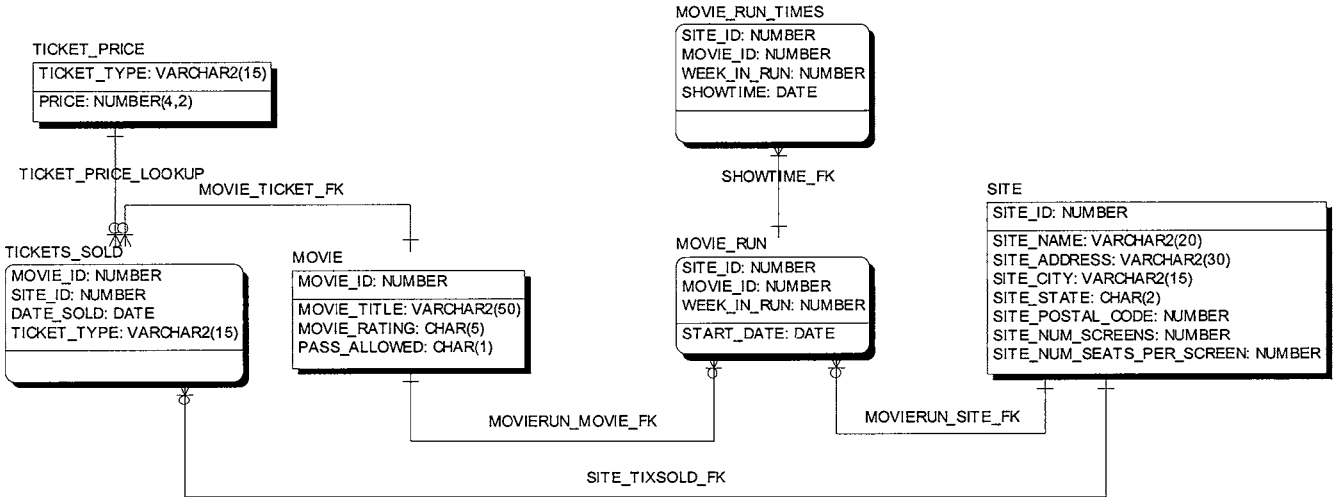
For example, for security reasons it has been determined that the finance manager will log all modifications to the EMPLOYEE_SALARY table for review. Granted, we could make use of a stored module (refer to the example above) and add the logging code into the stored module. However, we still have the loophole that the database could be broken into using the EMPLOYEE_SALARY table's login ID (or worse yet the database administrator's login!). If this were the case, the stored module would not be the only access point to the EMPLOYEE_SALARY table.

CONCLUSION

This chapter has given a brief introduction to the SQL language, including its historical and mathematical foundations. There are many good resources available online and in books, periodicals, and journals. Even though there is a standard for the SQL language, many vendors have added extensions and alterations that are often used to gain performance enhancements, which are very critical in Web apps. Complications in multiuser environments require that a locking analysis be conducted to detect and correct bottlenecks.

APPENDIX: SAMPLE SCHEMA AND DATA

Figure 1 depicts the tables that are used in this chapter and their relationships. The database supports the operation of a set of movie theaters (named SITES). Each site can run a MOVIE for a given number of weeks (this is stored in MOVIE_RUN). Because a movie can have different showtimes each week that it is being shown, this information is stored in the MOVIE_RUN_TIMES table. Of course, a movie cannot be shown without selling some tickets (TICKETS_SOLD). Because there are various charges for movies, this information is stored in the TICKET_PRICE table.



TICKET_PRICE

TICKET_TYPE	PRICE
KIDS	4.50
SENIORS	5.00
ADULT MAT	5.00
ADULTS	7.00

MOVIE

	MOVIE_ID	MOVIE_TITLE	MOVIE_RATING	PASS_ALLOWED
1	10001	THE BEST MAN IN GRASS CREEK	PG	Y
2	10002	A KNIGHT'S TALE	PG-13	N
3	10003	BRIDGET JONES"S DIARY	R	Y
4	10004	ALONG CAME A SPIDER	R	Y
5	10005	CROCODILE DUNDEE IN L.A.	PG	Y
6	10006	BLOW	R	Y
7	10007	DRIVEN	PG-13	Y
8	10008	EXIT WOUNDS	R	Y
9	10009	FREDDY GOT FINGERED	R	Y
10	10010	HEARTBREAKERS	PG-13	Y
11	10011	JOE DIRT	PG-13	Y
12	10012	FORSAKEN	R	Y
13	10013	MEMENTO	R	Y
14	10014	TOWN AND COUNTRY	R	Y
15	10015	THE MUMMY RETURNS	PG-13	Y
16	10016	ONE NIGHT AT MCCOOL"S	R	Y
17	10017	SPY KIDS	PG	Y
18	10018	THE TAILOR OF PANAMA	R	Y
19	10019	'CHOCOLAT'	PG-13	Y
20	10020	'AMORES PERROS'	R	Y

Figure 1: Relationship of tables used in this chapter.

GLOSSARY

Database Collection of relevant data organized using a specific scheme.

Relational database Collection of data that conforms to properties of the relational database model first defined by E. F. Codd.

Relational database management system (RDBMS) Software that manages a relational database.

SQL An industry-standard language for creating, updating, and querying a relational database.

Tuple Data object containing two or more components; usually refers to a record in a relational database.

CROSS REFERENCES

See *Data Mining in E-Commerce; Data Warehousing and Data Marts; Databases on the Web.*

REFERENCES

Celko, J. (1999). *Joe Celko's data & databases: Concepts in practice*. San Francisco: Morgan Kaufmann.

Chan, H. C., Tan, C. Y., & Wei, K. K. (1999, November 1). Three important determinants of user performance for database retrieval. *International Journal of Human–Computer Studie*s, 51, 895–918.

Connolly, T., & Begg, C. (2002). *Database systems*. Reading, MA: AddisonWesley.

Date, C. J. (1994, October). Moving forward with relational [interview by David Kalman]. *DBMS,* 62–75.

Eisenberg, A., & Melton, J. (2002). *SQL:1999* [formerly known as SQL3]. Retrieved April 24, 2002, from http://geochem.gsc.nrcan.gc.ca/miscellaneous_resources/sql1999.pdf

Elmasri, R., & Navathe, S. B. (1989). *Fundamentals of database systems*. New York: Benjamin/Cummings.

Freeon-line dictionary of computing (FOLDOC) (2002). Retrieved August 14, 2002, from http://wombat.doc.ic.ac.uk/foldoc/foldoc.cgi

Groff, J. R., & Weinberg, P. N. (1999). *SQL: The complete reference*. Berkeley, CA: Osborne McGraw–Hill.

Hursch, C. J., & Hursch, J. L. (1988). *SQL The structured query language*. Blue Ridge Summit, PA: Tab Professional and Reference Books.

Slazinski, E. D. (2001). Views—The 'other' database object. In D. Colton, S. Feather, M. Payne, & W. Tastle (Eds.). *Proceedings of the ISECON 2001 18th Annual Information Systems Education Conference* (p. 33). Foundation for Information Technology Education. Chicago: AITP.

SQL syntax diagrams (2002). Retrieved April 24, 2002 from http://www-dbs.inf.ethz.ch/~isk98/Syntax_Diagramme/SQL/

Supply Chain Management

Gerard J. Burke, *University of Florida*
Asoo J. Vakharia, *University of Florida*

INTRODUCTION

Supply chain management (SCM) is the art and science of creating and accentuating synergistic relationships among the trading members that constitute supply and distribution channels. Supply chain managers strive to deliver desired goods or services to the "right person," in the "right quantity," at the "right time," in the most effective and efficient manner. Usually this is achieved by negotiating or achieving a balance between conflicting objectives of customer satisfaction and cost-efficiencies. Each link in each supply chain represents an intersection where supply meets demand, and directing the product and information flows at these crossroads is at the core of SCM. The integral value proposition of SCM is as follows: Total performance of the entire chain is enhanced when all links in the chain are simultaneously optimized compared with the resulting total performance when each individual link is optimized separately. Obviously, coordination of the individual links in the chain is essential to achieve this objective. The Internet, and information technology in general, facilitate the integration of multitudes of channel enterprises into a seamless "inter"prise, leading to substitution of vertical integration with "virtual integration." Overcoming coordination roadblocks and creating incentives for collaboration among disparate channel members are some of the major current challenges in SCM.

How well the supply chain performs as a whole hinges on achieving fit between the nature of the products it supplies, the competitive strategies of the interacting firms, and the overall supply chain strategy. Planning also plays a key role in the success of supply chains. Decisions regarding facility location, manufacturing schedules, transportation routes and modes, and inventory levels and location are the basics that drive supply chains. These dimensions of tactical effectiveness are the sprockets that guide the chain downstream through its channel to end demand. Accurate and timely integrated information lubricate the chain for smooth operation. Information technologies allow supply chains to achieve better performance by providing visibility of the entire supply chain's status to its members, regardless of their position in the chain. The success of the collaborative forecasting, planning, and replenishment (CPFR) initiative illustrate the value of the internet to SCM (CPFR Committee, n.d.). A few of the most popular information tools or vehicles available to supply chains are enterprise resource planning (ERP) software and related planning applications, application service providers (ASP), online markets and auction mechanisms (business-to-business [B2B] commerce), and electronic customer relationship management (eCRM and business to consumer [B2C]).

A supply chain can be visualized as a network of firms servicing and being serviced by several other businesses, although it is conceptually easier to imagine a chain as a river, originating from a source, moving downstream, and terminating at a sink. The supply chain extends upstream to the sourcing of raw materials (backward integration) and downstream to the afterlife activities of the product, such as disposal, recycling, and remanufacturing (forward integration). Regardless of magnitude, all supply chains can be visualized as consisting of a sourcing stage, a manufacturing stage, and a distribution stage.

The supply chain operations reference model developed by the Supply Chain Council (n.d.) assumes that all processes at each of these stages are integral in all businesses. Each stage plays both a primary (usually physical transformation or service creation) and a dual (market mediator) role. This primary role depends on the strategy of the supply chain, which in turn, is a function of the serviced products' demand pattern (Fisher, 1997). The most strategic link is typically in the manufacturing or service creation stage, because it is positioned between suppliers and consumers. Depending on the structure of the chain (in terms of products and processes employed), power can shift from the supplier (e.g., monopolist supplier of key commodities such as oil), to the manufacturing (e.g., dominant producer of a unique product such as semiconductors), to the distribution (e.g., key distributor

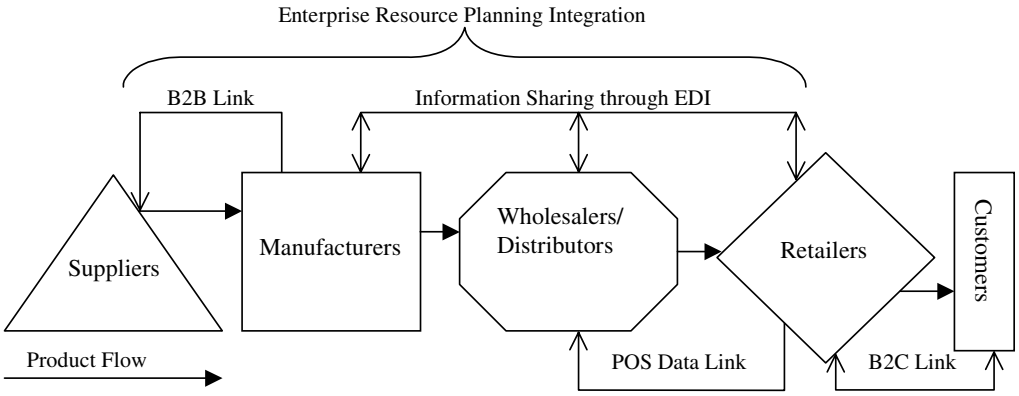


Figure 1: Example of an information-enabled supply chain.

of consumer items) stage in the supply chain. Obviously, power in the supply chain has a key bearing on the strategic positioning of each link in the chain.

The remainder of this chapter is organized as follows. In the next section, we describe the key strategic and tactical issues in SCM. This is followed by a discussion of mechanisms for coordinating stages in the supply chain. The third and final section focuses on describing the significant impact of information technology on SCM practices.

STRATEGIC AND TACTICAL ISSUES IN SCM

A holistic supply chain comprises multiple processes for suppliers, manufacturers, and distributors. Each process employs a distinct focus and a related dimension of excellence. Key issues in managing an entire supply chain relate to (a) analyzing product strategies, (b) network design and the related sourcing strategy employed, and (c) a strategic and tactical analysis of decisions in logistics, manufacturing, distribution, and after-sale service. It is our view that by analyzing product demand characteristics and the supply chain's capabilities, and crafting a fit between them, an individual manager can ensure that the specific process strategy employed does not create dissonance in the entire supply chain.

Product Strategies

SCM has evolved from process reengineering efforts to coordinate and integrate production planning at the factory level to initiatives that expand the scope of strategic fit beyond a single enterprise (Chopra & Meindl, 2001). Positive results from intra-functional efforts extended the SCM philosophy throughout the enterprise. Process improvements at the firm level highlighted the need for suppliers and customers of supply chain managed firms to adopt the SCM philosophy. A supply chain is only as strong as its weakest link. How the chain defines strength is at the core of a supply chain's strategy and, therefore, its design. Is strength anchored in efficiency, responsiveness, or both?

Achieving a tight fit between the competitive strategies of supply chain members and the supply chain itself is gained by evaluating the characteristics of the products serviced by the chain. "The root cause of the problems plaguing many supply chains is a mismatch between the type of product and the type of supply chain" (Fisher, 1997, p.105). Critical product attributes are (a) the demand pattern, (b) the life cycle, (c) the variety of offerings, and (d) the product delivery strategy. Fisher (1997) categorized a product as being either functional (basic, predictable, long-lived, low profit margin) or innovative (differentiated, volatile, short-lived, high profit margin). Furthermore, using the product life-cycle argument, innovative products (if successful) will eventually evolve to become functional products. The types of supply chain needed to service these two categories of products effectively are distinct. An efficient or low-cost supply chain is more appropriate for a functional product, whereas a responsive or customer attuned supply chain fits an innovative product better. Obviously, a spectrum of chain varieties exists between the end points of responsiveness and efficiency, hence most tailored supply chains are hybrids that target responsiveness requirements for each product serviced while exploiting commonalities in servicing all products to gain economies of scope. Thus, the strategic position of a supply chain balances customer satisfaction demands and the firm's need for cost minimization.

Information technologies enable both efficient and responsive supply chains because they have the potential to provide immediate and accurate demand and order status information. Efficiency gains via information technologies are gleaned from decreased transactional costs resulting from order automation and easier access to information needed by chain members. Likewise, responsiveness gains can be obtained by a quicker response to customer orders. Hence, in practice, it seems to have become standard for all supply chains to use some form of information technology to enable not only a more efficient physical flow of their products but also to simultaneously improve their market mediation capability.

Network Design

Network decisions in a supply chain focus on facility function, facility location, capacity, and sourcing and distribution (Chopra & Meindl, 2001). In general, network design

determines the supply chain's tactical structure. The significant capital investments required in building such a structure indicate the relative long run or strategic importance of network decisions.

Facility Function

How will each network investment facilitate the supply chain strategy? Consider a manufacturing plant. If the plant is set up to produce only a specific product type, the chain will be more efficient, but less flexible than it would be if the plant produced multiple product types. A supply chain providing innovative products will likely perform better with flexible manufacturing facilities.

Facility Location

What locations should be chosen for facilities? A good decision in this dimension is essential because the cost ramifications of a suboptimal location could be substantial. Shutting down or moving a facility is significant not only in terms of financial resources but also in terms of the impact on employees and communities. Other factors that should be considered are the available infrastructure for physical and information transportation, flexibility of production technologies employed, external or macroeconomic influences, political stability, location of competitors, availability of required labor and materials, and the logistics costs contingent on site selection.

Capacity

Depending on the expected level of output for a facility, capacity allocations should be made so that idle time is minimal. Underutilization results in lower return on investment and is sure to get the attention of company executives. On the other hand, underallocating capacity (or large utilizations) will create a bottleneck or constricted link in the supply chain. This will result in unsatisfied demand and lost sales or increased costs as a result of satisfying demand from a nonoptimal location. The capacity allocation decision is a relatively long-term commitment, which becomes more significant as sophistication and price of the production technology increases.

Supply and Distribution

Who will serve our needs, and whose needs will we serve? This is a recurring question. Decisions regarding the suppliers to a facility and the demand to be satisfied by a facility determine the costs of material inputs, inventory, and delivery. Therefore, as forces driving supply or demand (or both) change, this decision must be reconsidered. The objective here is typically to match suppliers and markets to facilities to minimize not only the systemwide costs but also the customer responsiveness of the supply chain. In general, these criteria are often orthogonal, implying that the sourcing and distribution decisions require a multiobjective focus with some prioritization of the cost and responsiveness aspects.

Each of these network design decisions are not made in isolation since there is a need to prioritize and coordinate their combined impact on the tactical operations of the supply chain. They jointly determine the structure of the supply chain, and it is within this structure that tactical strategies are implemented to reinforce the overall strategy of the entire chain. Once network design has been finalized, the next key decision within the supply chain focuses on sourcing strategies.

Sourcing Strategies

A primary driver of a firm's SCM success is an effective sourcing strategy. The firm's ability to deliver its goods and services to customers in a timely manner hinges on obtaining the appropriate resources from the firm's suppliers. Because manufacturing firms typically spend more than 50% of earned revenue on purchased materials, the costs of disruptions to production due to supply inadequacies are especially significant. Furthermore, a firm's financial and strategic success is fused to its supply base.

The nature of a firm's connection to its suppliers is evident in its sourcing strategy and is characterized by three key interrelated decisions: (a) how many suppliers to order from, (b) what criteria to use for choosing an appropriate set of suppliers, and (c) the quantity of goods to order from each supplier.

Single sourcing strategies seek to build partnerships between buyers and suppliers to foster cooperation and achieve benefits for both players. With the adoption of just-in-time inventory policies, supplier alliances with varying degrees of coordination have shifted supply relations toward single sourcing to streamline the supply network. At the strategic level, single sourcing contradicts portfolio theory. By not diversifying, a firm is assuming greater risk. Therefore, tactical single sourcing benefits need to justify the riskiness inherent in this relationship. One method of alleviating the risks of single sourcing is to ensure that suppliers develop contingency plans for materials in case of unforeseen circumstances. In fact, suppliers dealing with IBM are required to provide details of such contingency plans before the company will enter into purchasing contracts with them.

The obvious benefits of single sourcing for the firm are quantity discounts from order consolidation, reduced order lead times, and logistical cost reductions as a result of a scaled down supplier base. The ordering costs advantage of single sourcing is diminishing, however, because of the proliferation of Internet procurement tools, which tend to reduce ordering costs and streamline purchasing processes.

In contrast, a larger supply base possesses greater upside volume flexibility through the summation of supplier capacities. Strategically, a manufacturer's leverage is kept intact when the firm diversifies its total requirements among multiple sources. Additionally, alternative sources hedge the risk of creating a monopolistic (sole source) supplier, and the risk of a supplier integrating forward to compete with the buying firm directly. Online exchanges and marketplaces also provide multiple sourcing benefits by automating the cumbersome tasks associated with multiple supplier dealings.

In concert with decisions regarding the number of suppliers for a product, a firm must develop an appropriate set of criteria to determine a given supplier's abilities to satisfy the firm's requirements. In practice, this is evaluated using scoring models which incorporate quantifiable

and qualitative factors related to quality, quantity, delivery, and price. Although the supplier's price may be the most important criteria for generic or commodity-type goods, other dimensions incorporated in these models are probably more or equally important for innovative products. Thus, supplier selection is not simply a matter of satisfying capacity and price requirements, but also needs to integrate supplier capabilities in terms of quality and delivery. It should be obvious that it is the collective suppliers' capabilities that can enable or limit supply chain performance at its inception.

Once an appropriate set of suppliers has been identified, the firm must determine a suitable order quantity for each vendor. Because the overall demand for the firm's goods are typically uncertain, newsboy-type models can be successfully used in this context. Essentially, these models take into account the cost of inventory (overage cost) and the cost of unsatisfied demand (underage cost) to determine an appropriate order quantity. Ultimately, decisions regarding the number of suppliers from which to purchase, supplier selection, and order quantity allocation among selected suppliers should support the supply chain strategy's focal purpose of being efficiently responsive or responsively efficient. Sourcing decisions are one of the primary drivers of transportation and logistics-related issues discussed in the following section.

Transportation and Logistics

Transportation decisions affect product flow not only between supply chain members but also to the market place. In most supply networks, transportation costs account for a significant portion of total supply chain cost. Transportation decisions are mainly tactical, however. In determining the mode(s) and route(s) to employ through the supply chain, transportation decisions seek to strike a balance between efficiency and responsiveness so as to reinforce the strategic position of the supply chain. For example, an innovative product's typically short life cycle may warrant expensive air freight speed for a portion or all of its movement through the chain, whereas a commodity is generally transported by slow but relatively economical water or rail freight. Shipping via truck also is used frequently. Trucking is more responsive and more expensive than rail and less responsive and less expensive than air. Most supply chains employ an intermodal strategy (e.g., raw materials are transported by rail or ship, components by truck, and finished goods by air).

A supply chain's transportation network decisions are inextricably linked to its network design decisions. Transportation network design choices drive routing decisions in the supply network. The major decisions are whether to ship directly to buyers or to a distribution center and whether a routing scheme is needed. A direct shipping network ships products directly from each supplier to each buyer. A "milk run" routed direct shipping network employs a vehicle to ship either from multiple suppliers to a single buyer or from a single supplier to multiple buyers. A network with distribution centers ships from suppliers to a warehouse or distribution facility. From this facility, buyers' orders are picked from the distribution center's inventory and shipped to the buyers. This design can also employ milk runs from suppliers to the distribution center and from the distribution center to the buyers. Through the supply network, combinations and variants of these designs that best suit the nature of the product provide locomotion from supplier to buyer.

B2C transactions enable end consumers to demand home delivery, creating a larger number of smaller orders. This trend has enhanced the role of package carriers such as FedEx, United Parcel Service, and the U.S. Postal Service in transporting consumer goods. Additionally, bar coding and global positioning systems (GPS) provide accurate and timely information of shipments enabling both buyers and suppliers to make better decisions related to goods in transit. Technological advancements of the past decade enable consumers to demand better responsiveness from retailers. The resulting loss of some scale economies in shipping, have been offset by shipping cost decreases due to information-technology-enabled third-party shippers that more efficiently aggregate small shipping orders from several suppliers. As consumers' expectations regarding merchandise availability and delivery become more instantaneous, the role of a supply chain's transportation network in overall performance expands. This provides the link to operations and manufacturing issues because the role of information technology has forced these processes in a supply chain to be more responsive.

Operations and Manufacturing

"A transformation network links production facilities conducting work-in-process inventories through the supply chain" (Erenguc, Simpson, & Vakharia, 1999, p. 224). Suppliers linked to manufacturers linked to distribution systems can be viewed as a transformation network hinging on the manufacturer. Transforming supplies begins at the receiving stations of manufacturers. The configuration of manufacturing facilities and locations of transformation processes are determined by plant-level design decisions. The manufacturing process employed at a specific plant largely drives the decisions. An assemble-to-order plant may have little investment in production but a great deal of investment in storage. A make-to-stock facility may have little or no investment in component inventory but a great deal of investment in holding finished goods inventory. A make-to-order factory may have significant investment in components and production facilities, and no finished goods investment. A product's final form can also take shape closer to the end consumer. To keep finished goods inventory costs as low as possible and to better match end demand, a supply chain may employ postponement to delay customizing end products.

Major design decisions such as facility configuration and transformation processes are considered long-term decisions, which constrain the short- to mid-term decisions addressed in a plant's aggregate plan, a general production plan that encompasses a specific planning horizon. Information required to develop an effective aggregate plan include accurate demand forecasts, reliable supply delivery schedules, and the cost trade-offs between production and inventory. Each supply chain member develops an aggregate plan to guide short-term operational

decisions. To ensure that these individual plans support each other, the planning process must be coordinated. The degree and scope of coordination will depend on the economics of collaborative planning versus the costs of undersupply and oversupply. It is likely unnecessary and impossible to involve every supply chain member in an aggregate plan for the entire supply chain; however, a manufacturer should definitely involve major suppliers and buyers in aggregate planning. Whether this planning information trickles to other supply chain members (a key for the success of integrated supply chain management) will depend on the coordination capabilities of successive layers of members emanating from a collaborative planning center, which is often the major manufacturer.

The strategy employed to execute the aggregate plan is a function of the information inputs into the aggregate plan. It is vital that these inputs be as accurate as possible throughout the entire supply chain. Integrated planning in a supply chain requires its members to share information. The initiator of integrated planning is typically the major manufacturer. To understand why, we must understand the dynamics of distribution.

Distribution Channels

To anticipate the quantity of product to produce, a manufacturer must compile demand forecasts from downstream supply chain members. Forecasting accuracy is paramount because it is the basis for effective and efficient management of supply chains. The root challenge of SCM is to minimize costs and maintain flexibility in the face of uncertain demand. This is accomplished through capacity and inventory management. Similarly, marketers attempt to maximize revenues through demand management practices of pricing and promotion. Therefore, it is vital that marketing and operations departments collaborate on forecasts and share harmonious incentive structures. The degree of coordination among order acquisition, supply acquisition, and production process directly affects how smoothly a firm operates. Likewise, the coordination level of buyers, suppliers, and producers directly affects how smoothly the supply chain operates. More specifically, accurate information flows between channel members are essential to SCM.

A distribution channel is typically composed of a manufacturer, a wholesaler, a distributor, and a retailer. The "bull-whip effect" is a classic illustration of dysfunction in such a channel due to the lack of information sharing. This effect is characterized by increasing variability in orders as the orders are transferred from the retailer upstream to the distributor, then to the wholesaler, and finally to the manufacturer. Distorted demand information induces amplifications in variance as orders flow upstream. Therefore, the manufacturer bears the greatest degree of order variability. It is for this reason that manufacturers often initiate collaborative efforts with downstream channel members.

Lee, Padmanabhan, and Whang (1997) analyze four sources of the bull-whip effect that correspond to respective channel practices or market conditions. The first two causes are a direct consequence of channel practices, whereas the latter two causes are more market-driven. The first source, demand signal processing, is largely due to the use of past demand information to modify demand forecasts. Each channel member modifies her or his forecasts and resulting orders in isolation. These multiple forecasts blur end demand. This problem is exacerbated as lead time lengthens. Current practices employed to remedy this source of information distortion include contractual agreements to provide point of sale data from retailers to manufacturers, vendor managed inventory to centralize ordering decisions, and quick response manufacturing to decrease lead times for order fulfillment. The second source of information distortion, order batching, arises mainly from periodic review ordering practices and processing costs of placing orders. Without specified ordering times, the timing of order placement for several vendors may coincide as a result of fiscal period delineations. Additionally, a buyer may accumulate orders to minimize high costs of ordering and shipping. Measures that can alleviate these causes are automated and fixed-time ordering, electronic data interchange (EDI), and the use of third-party logistics providers to offset less than truckload diseconomies. These first two causes of information distortion generally result from each channel member individually optimizing inventory decisions.

Price fluctuations, the third cause of the bull-whip effect, result from marketing efforts such as trade promotions to generate increases in sales volumes. Wholesale price discounts result in forward buying by retailers. The lower price motivates retailers to stock up on product for not only the current but also future periods. This strategy results in uneven production schedules for manufacturers and excess inventory carrying costs for retailers. The push for everyday low prices is an effort to do away with trade-promotion-induced order variability. The final cause of the bullwhip effect occurs when there exists a perceived shortage of product supply. If the supplier adopts a rationing scheme that is proportional to the quantity ordered, buyers will simply inflate their orders to ensure receipt of their true requirements. This type of gaming can be avoided by rationing based on historical market share of buyers and information sharing between buyers and suppliers to prevent supply shortages.

This concludes our discussion of key strategic and tactical issues in managing supply chains. In general, it is obvious that chain members are motivated to minimize information distortion and resultant order variability. It is also clear that the remedies to these problems are rooted in information sharing upstream through the supply chain. Coordination mechanisms facilitating information sharing in a supply chain are described in the next section.

SUPPLY CHAIN COORDINATION

In a supply chain, coordination occurs when the constituents act in unison for the betterment of the supply chain as a whole rather than their own link. Thus, it is likely that on an individual basis, a chain member may stand to suffer a "loss" associated with a coordinated decision. In economics, this is the classical principal-agent dilemma. Supply chain coordination is fraught with

impediments stemming from human nature and technological limitations. On the other hand, in addition to "bull-whip" remedies, mechanisms do exist to promote coordination within the supply chain, and these are described next.

Incentive Impediments

Lee and Whang (1999) discuss incentive problems that prevent coordination in a supply chain with a decentralized decision structure. In this type of system, each site manager makes decisions to optimize his or her personal benefit. This type of incentive misalignment can exist within and between functional areas of a site (firm) and also between sites (firms) in the supply chain. For instance, a marketing manager's objective may be revenue maximization, and he or she may attempt to generate sales with trade promotions to induce forward buying. Meanwhile, a manufacturing manager may have a conflicting objective, such as minimizing production variability or level utilization. As illustrated through the bull-whip effect, trade promotions generally increase variability in production.

These types of conflicts of interest can be addressed by corporate operating guidelines or rules. For instance, manufacturing representatives should be evaluated over a rolling horizon on average sales of their products by their customers instead of to their customers. This performance measurement scheme removes the incentive to foster forward buying and reduces cycle inventories. Although, this type of approach works well for allaying conflicts within a firm, to mitigate incentive misalignment between firms, contractual relationships are formed. From a supply chain perspective, two obvious conflicts of interest exist between upstream managers and retail managers. The upstream managers covet end demand information that retailers own and which, in isolation, they have no incentive to share. Likewise, if retailers are the only site managers penalized for stock-outs, upstream sites have no incentive to carry safety stock even though they typically incur lower inventory carrying costs than retailers. To reconcile conflicting interests, channel members resort to manipulating performance requirements embedded in contractual agreements.

Fostering Trust between Partners

Contractual agreements can also be used to develop a foundation of trust between trading partners. A simple contractual relationship is far from a collaborative relationship, however. Firms' inability to collaborate has often been blamed on technological limitations of information management. Interestingly, Gallagher (2001) commented that during a spring 2001 Supply Chain Council executive retreat, that attendees cited a "lack of trusted relationships between key supply chain participants" as the main obstacle to effective collaboration. Information sharing is no longer constrained by information technologies; it is held back by the impersonal nature of automation. Collaboration, especially information sharing, is limited by shallow business relationships. It is for this reason that collaborative supply chains seek to identify channel members that add significant value and develop richer relationships with those firms. Higher levels of trust can be achieved through elaborate and exhaustive contingency contracts or by interacting over time in a consistent mutually beneficial manner. These relationships can be sustained if both parties are mutually interdependent and gain mutual benefit from the partnership. Whether the investment in strategic alliances is justified will depend largely on the supply chain's strategy.

Incentives from Terms of Sale

Supply chain coordination can be accomplished through a quantity discount schedule for commodity type products. Because prices for these products are set by the market, a lot size quantity discount scheme can coordinate the supply chain. This holds as long as the increased cycle inventory costs are less than the benefits of the quantity discount scheme. This mechanism is well suited for supply chains with an efficiency driven strategy associated with commodities. A related approach for incentivizing suppliers is to enter into long-term blanket orders with multiple order releases within the planning horizon.

If a seller has market power through a patent, copyright, and so on, channel coordination can be achieved through volume-based quantity discounts. The key difference between this scheme and the lot size quantity discount is that the volume-based discount is based on the rate of purchase instead of the amount purchased per order. An example of a volume-based quantity discount is a two-part tariff. With this arrangement, the seller charges an initial fee to cover his profit, and then unit prices are set to cover production costs.

Another way sellers can induce buyers to purchase greater quantities is to have a returns policy. In offering a buyback contract, a seller stipulates a wholesale price per unit as well as a buyback price. The optimal order quantity for the buyer then rises as a result of a guaranteed salvage value for unsold products. In situations in which the costs of returns are high relative to the salvage value of the product, sellers may coordinate the supply chain through quantity flexibility contracts. This arrangement allows buyers to modify their order quantity after observing their respective demand. Rather than committing in advance of observing demand to a specific order quantity, the buyer commits to a minimum order quantity, and the seller commits to providing a maximum quantity. Total supply chain profits can increase as a result of this contractual agreement. By manipulating the terms of sale, sellers can induce buyers to purchase in greater quantities, which ensures sellers greater profitability. What is somewhat surprising is that buyers, too, can gain increased profits if these terms of sale are appropriately set.

Supply chain coordination can be achieved if incentives are aligned throughout the supply chain via contracts or consistent performance measures, information passes accurately through the supply chain, operations perform optimally, and the requisite level of trust exists within the supply chain. It is therefore understandable that until recently, achieving channel coordination or optimizing total supply chain performance seemed impossible.

With advances in information technologies over the past decade, however, this pipe dream may actually become reality.

INFORMATION TECHNOLOGY AND SCM

At the overall level, the broadest impact of information technology on SCM is by enabling tighter connectivity between supply chain members. The time and distance gaps that previously hampered information flows and supply chain responsiveness are all but eradicated by the Internet's relative immediacy of access to existing information. Keskinocak and Tayur (2001, p. 71) noted that the Internet's rapid communication, innovative trading spaces, accessibility of new distribution channels, and facility for collaboration encourages supply chain managers the rationale for envisioning virtual integration as a reality. Obviously, sharing of accurate information is essential for collaboration. To this end, trading partners must have a means to exchange information. The explosion of ERP implementations in the 1990s "dramatically improved the quantity and quality of data" that could be used for effective SCM (Sodhi, 2001, p. 56). Current information technologies such as enterprise applications, on-line auctions, e-markets, and web-based services provide gears that supply chain managers can shift to tailor the sequence of paths taken by the chain to best move its product(s).

Enabling Collaboration

ERP systems, offered by firms such as SAP and Oracle, provide the basic functionality of visibility and tracking on which tighter connectivity is executed. For example, through Web interfaces, customers and suppliers can place orders to the firm or provide replenishment information for the firm. The ERP system pulls this information from a database and compiles an enterprisewide view that provides order status information for customers, generates shipping orders to suppliers, and inventory positions to APS applications offered by SAP, Oracle, i2 Technologies, Manugistics, and others.

APS provides planning and execution functionality, and these products have become more focused and refined to keep pace with empowered customers seeking efficiencies through collaboration (Harreld, 2001). More specialized SCM applications for supply management generally are execution or planning focused. For example Sears, Roebuck & Company and its major appliance suppliers use a SeeCommerce execution focused application to manage supply chain performance. This application pulls data from transactional purchasing and quality assurance systems to generate real-time performance metrics that are visible to each supplier. This allows Sears and its appliance vendors to address problems proactively before they become unwieldy. Alternatively, a planning focused SCM application by SynQuest is used by Ford Motor Company to optimize auto part delivery costs in a just-in-time environment. This software models Ford's myriad of inbound logistics and evaluates several variables to determine the best movement of inbound parts. As a result of this implementation, missing parts errors have improved 100-fold.

Application service providers (ASP) play a key role in information-enabled supply chain coordination and collaboration by providing Web-based focused solutions for tactical planning. The standardization of data largely due to the adoption of XML (extensible markup language) allows disparate ERP and ASP applications to transact. ASPs provide small supply chain members with small budgets an avenue to tighter connectivity. In additional to B2B portal provision, ASPs provide value to their customers by providing software applications, access to hardware, or consulting services to allow customers to outsource some or all of their IT functions effectively. Standardization efforts through the CPFR Committee strive to foster effective SCM via the internet. Furthermore, interoperability of applications across the supply chain appears more viable with Web services platforms such as Microsoft.NET.

Auctions and Exchanges

Major ERP vendors also provide products that enable firms to create online auctions. The time and financial resources spent developing and organizing an auction are significant overhead expenses. Therefore, industry partnerships or third parties host many auctions and e-markets. B2B (e.g., Covisint) and B2C (e.g., Onsale) exchanges are dramatically changing the manner in which supply chain activities are being structured. The underlying structure of these exchanges is built around an auction mechanism. In general, an auction can be viewed as a way for sellers to obtain information on the reservation prices of potential customers in the market for a particular product. The implementation of auctions on the Web provides a guarantee for consumers in terms of price efficiencies and simultaneously helps to match buyers and sellers.

Electronic End Demand Fulfillment

With ever-increasing consumer demands for customization and competition, and ever-decreasing product life cycles, SCM has gravitated toward a sense and respond approach. Sodhi (2001) provided the following examples of Internet enabled demand fulfillment processes:

Design

The data regarding end-consumer-preferred product attributes can be gathered quickly and efficiently through Web surveys. Designers can also collaborate and plan product launch particulars via the Internet to speed up time to market.

Customer-Relationship Management

Data collected from navigational paths of individual Web site visitors can be used to predict purchasing behavior, which can lead to better forecasts and inventory decisions for e-businesses. For example, collaborative filtering software compiles profiles based on customers browsing or purchasing behavior to fit them into a market segment. Based on this real-time classification, advertisements are displayed to entice initial or further purchases. Collaborative filtering is one method of personalization. Personalization attempts to modify Web site offerings to match

the interests of individuals. Another personalization tool is "clickstream" analysis.

Self-Service

Web-based package tracking provided by United Parcel Service and FedEx is an example of the Internet providing better service to customers at a lower cost per transaction to the seller. Other online applications read e-mail from customers to sort or even respond to e-mail, providing more timely feedback to customers.

Disintermediation

The current state of information technology also expands the options available for distributing products. Manufacturers that were once reliant on wholesalers, distributors, and retailers can more effectively market directly to end customers. The success of Dell's direct model, which leverages Internet technologies, has encouraged other firms to pursue direct selling. Nonetheless, although one can eliminate a link in a supply chain, the functions that link performs cannot be eliminated.

SUMMARY

At the core of SCM is the optimal coordination of activities at and between individual supply chain members. We have outlined how supply chain members individually and collectively must support and enhance the characteristics of the products serviced by the chain. Friction between product type and the type of supply network servicing the product may be the underlying cause of many supply chain problems. Given that product and supply chain strategies are aligned, effective execution of the SC's strategy is built on operational foundations including overall network design, sourcing, logistics, transformation processes, and distribution channels. These basic activities work in concert to balance requirements of quality or service levels with economic realities of cost containment. Supply chain managers invariably rely on information technology to facilitate more efficient physical flow of their products, and foster better market mediation throughout the chain.

Coordinating the activities of the links in the supply chain to act in the best interest of the entire chain is a major challenge. The information technology required to streamline transactional processes and enable integrated decision making throughout the supply chain exists. Therefore, the key to accomplishing coordination is establishing and maintaining trust between trading partners. SCM seeks to elevate the supply network's performance with technology while enhancing business relationships between its constituencies. The science of SCM demonstrates that the supply network's sum is greater than sum of its parts, while persuading decentralized members of the network to act in the best interest of the network overall is probably more of an art than a science.

GLOSSARY

Application service provider (ASP) A remote provider that hosts software and provides access for a periodic rental fee.

Advanced planning and scheduling (APS) Applications employed to optimize production schedules or supply chain activities.

Business-to-business (B2B) Electronic business conducted between two firms.

Business-to-consumer (B2C) Electronic business conducted between a firm and a consumer.

Collaborative planning forecasting and replenishment (CPFR) Efforts to coordinate supply chain members through point-of-sale data sharing and joint planning.

Distribution channel Supply chain members involved in moving end products to consumers.

Electronic customer relationship management (eCRM) Applications designed to automate demand satisfaction.

Enterprise Resource Planning (ERP) Information systems that keep track of transactional data to provide decision makers with better information.

Newsboy model Inventory model to determine the optimal balance between stock-out and inventory holding costs.

Postponement Delaying product customization until closer to the point of sale.

Vendor managed inventory A suppliers' making replenishment decisions based on retailer sales data for products at retail sites.

Extensible markup language (XML) Language designed to describe data that is key to efficient Web-based data transfer and manipulation.

CROSS REFERENCES

See *Application Service Providers (ASPs); Developing and Maintaining Supply Chain Relationships; Electronic Procurement; Enterprise Resource Planning (ERP); International Supply Chain Management; Supply Chain Management and the Internet; Supply Chain Management Technologies*.

REFERENCES

Chopra, S., & Meindl, P. (2001). *Supply chain management: Strategy, planning and operation*. Upper Saddle River, NJ: Prentice-Hall.

Collaborative Planning, Forecasting and Replenishment Committee Web site (n.d.). Retrieved March 10, 2002, from http://www.cpfr.org/

Erenguc, S. S., Simpson, N. C., & Vakharia, A. J. (1999). Integrated production/distribution planning in supply chains: An invited review. *European Journal of Operational Research, 115,* 219–236.

Fisher, M. L. (1997). What is the right supply chain for your product? *Harvard Business Review, 75,* 105–117.

Gallagher, P. (2001). Where's the trust? Retrieved April 25, 2002, from http://www.e-insite.net/electronicnews/index.asp?layout = article&articleId = CA186834

Harreld, H. (2001). Supply chain collaboration. Retrieved April 25, 2002, from http://staging.infoworld.com/articles/fe/xml/01/12/24/011224fesccollab.xml?Template = /storypages/ctozone_story.html&Rsc = 2

Keskinocak, P., & Tayur, S. (2001). Quantitative analysis for Internet-enabled supply chains. *Interfaces, 31,* 70–89.

Lee, H. L., Padmanabhan, V., & Whang, S. (1997). Information distortion in a supply chain: The bullwhip effect. *Management Science, 43*, 546–558.

Lee, H. L., & Whang, S. (1999). Decentralized multi-echeleon supply chains: Incentives and information. *Management Science, 45*, 633–640.

Sodhi, M. S. (2001). Applications and opportunities for operations research in Internet-enabled supply chains and electronic marketplaces. *Interfaces, 31*, 56–69.

Supply Chain Council Web site. (n.d.). Retrieved March 10, 2002, from http://www.supply-chain.org/

Supply Chain Management and the Internet

Thomas D. Lairson, *Rollins College*

INTRODUCTION

The Internet creates a new environment for exchanging information and conducting business transactions. More than ever possible before, the Internet increases the quantity and expands the richness of information in real time to a much wider set of participants and thereby raises dramatically the value of information in supply chain management. The Internet also increases transparency, which is the ability to "see across the supply chain," through the enhanced capacity to obtain, distribute, and create information across distances, and does so at a cost that has been decreasing by 35% per year.[1] This ability changes the management of supply chains by expanding capabilities and extending the scope of management across the supply chain, and contributes to reductions in cost and improved service. As a consequence, the strategic calculations of firms must now incorporate supply chain operations and new business models that concentrate on reaping the benefits of an Internet-based supply chain. This places a tremendous premium on creating, sharing, and using information throughout the supply chain. Where once firms in a supply chain had little choice but to operate mostly alone in an information vacuum, now they must operate in a collaborative environment with an abundance of information. Trade-offs that once seemed immutable are now significantly modified by new Internet capabilities.

THE GLOBAL BUSINESS REVOLUTION AND SUPPLY CHAINS

The Internet has emerged as a new medium of business at a time of revolutionary changes in global business, and it will reinforce and accelerate these changes. Many traditional business problems and issues will continue into the Internet era, but with new technological options and management challenges. The business environment began to change in the 1970s, as a result of the globalization of production, trade, and capital flows (Nolan, 2001). Making this possible were the familiar forces of falling trade barriers, liberalization of capital movements, and declining transportation, communication, and information processing costs. The result was a transformation of the production of goods, with a focus on cost reduction through lean manufacturing, shortened product cycles and more flexible manufacturing capabilities, and increasing pressure for greater product variety and customization. The new competition made many firms rethink their capabilities and shed secondary operations so as to focus on "core capabilities." The most competitive firms were able to develop a global brand and amass the product technology, R&D, financial strength, information technology, and human resources necessary to develop and manufacture products through globally constructed supply chains. These "core system integrators" were successful in assembling the most globally competitive suppliers—usually those with the ability to invest in the needed R&D and information technology—and this forced second-tier suppliers to make similar investments needed to participate in the system. The consequence was a geographical dispersion of production and a much finer division of labor within supply chains, with ever more differentiated product components located in ever wider geographical areas. This global value chain typically is tied together by the "system integrator," a firm that is especially effective in gathering and distributing information on a global scale.

These global supply networks have restructured the competitive position of firms and the development opportunities of nations. Supply chains have become a central source of competitive strength, operating as the main ingredient for the development of a system of lean, flexible production, customization, and rapid, accurate distribution. Firms have been forced to upgrade the efficiency of their supply chains, and this process raises the value of information dramatically. Many of these changes had occurred by the mid-1990s, when the Internet emerged

[1] This calculation is derived from Moore's Law, which describes the rate of increase in the number of transistors on a fixed size of semiconductor material. Gordon Moore predicted, correctly, that this number would double every 18 months. The consequence is that the processing power of a computer doubles every 18 months, at a constant to falling cost. The result is that the cost of information creation and distribution falls by roughly 35% per year and has done so for more than 30 years (Woodall, 2000).

as a commercial instrument. Building on already developed technologies such as electronic data interchange (EDI) and satellite communication, the Internet has become a central technology for advancing the revolutionary changes occurring in the nature and location of production.

HOW DOES THE INTERNET AFFECT SUPPLY CHAIN MANAGEMENT?

The management of a supply chain in the era of globalization, lean and flexible manufacturing, and mass customization presents managers with a formidable set of tasks. The need for lower costs, greater speed, more flexibility, and increased service generates a series of optimization problems compounded by a complex array of trade-offs. Because the Internet is a cost-effective and near-ubiquitous medium of information exchange, it expands the opportunities for coping with these difficulties. The most direct effect of the Internet is to create new opportunities to improve the efficiency and effectiveness of the operation of the supply chain. This is because of the cost-effective capacity to generate visibility across all aspects of the supply chain, including point-of-sale information, manufacturing schedules, vendor stocks, customer inventories, demand patterns, sales/marketing initiatives, and carrier schedules. However, achieving these gains requires many management decisions and organizational changes, most of which are focused on recognizing the value of collaboration among members of the supply chain and designing a system to facilitate this collaboration. At the same time, barriers to collaboration arise from the different interests and needs of the members of the supply chain, and because the benefits are not equally available to all members. Collaboration is also difficult to achieve because it requires a rethinking of the nature of the business enterprise and the relationships among external suppliers, core business operations, and customers.

The application of the Internet to the management of supply chains is one of the most important forms of e-business. Supply chain management is the organization, design, optimization, and utilization of the business processes and physical and information networks that link raw materials to the delivery of end products to customers. The supply chain is a system of production, assembly, exchange, information flows, financial flows, transactions, physical movement, and coordination. The business processes involved include relationships with customers, fulfillment of orders, payment, management of demand, procurement of materials and supplies, manufacturing coordination, and the logistics of moving materials and products (Lambert, 2001).

However, what is e-business? By one definition, it is the "marketing, buying, selling, delivering, servicing, and paying for products across (nonproprietary) networks to link an enterprise with its prospects, customers, agents, suppliers, competitors, allies, and complementors" (Weill & Vitale, 2001, p. 5). Others have offered a view emphasizing e-business applications to the supply chain: "planning and execution of the front-end and back-end operations in a supply chain using the Internet" (Lee & Whang, 2001a, p. 2).

The application of the Internet to supply chain management involves developing the capacity for greater integration, for new forms of collaboration, and for using the new information systems to redesign business practices.

We will approach the question of e-business from the perspective of the extended, real-time enterprise (Siegele, 2002) and consider not only the internal integration of the supply chain but also how linking the supply chain to customers and strategic partners solves problems and expands opportunities. The extended, real-time enterprise is a core concept for understanding the operation of e-business. It involves the development of an integrated information system linking together customers, the firm and its operations, strategic partners, and the relevant supply chains. Information about orders, operations, suppliers, and logistics is available in real-time and permits new forms of management and relationships with customers and suppliers.

Opportunities for Gains

The greatest barrier to improvement in the operation of supply chains is the segmentation and separation of the various elements of the supply chain. The different units of a supply chain—suppliers, manufacturers, distributors, logistics, and retailers—have typically operated without the information integration, synchronization, and coordination needed to improve operations. Even when firms in a supply chain have wanted to improve integration, because information has traditionally been expensive to generate and distribute, they have been forced to operate in relative isolation. This contributes to higher inventory levels, difficulties in responding to customers in a timely manner, ineffective use of resources, problems in developing new products, limited knowledge of customer demand, and lower profits. Because the Internet generates more and better information in real time and shares that information in a cost-effective way across the supply chain, it offers efficiency gains from speed, cost, flexibility, and expanded service. Put another way, these capabilities result in reduced segmentation and separation and greater integration of the supply chain through information sharing.

The Internet, with the capacity to generate and distribute information at low cost, provides significant opportunities for integration of the elements of the supply chain. As Figure 1 suggests, information flows in an Internet-enabled supply chain are much more complex than in a tradition supply chain. To be the most effective, information flows must be continuous and must link all parties in the chain, so that each firm can see what is happening throughout the system. Thus, the potential benefits from the Internet require much greater collaboration, since information generated at each point of the supply chain must be shared, and decisions regarding this information must be based upon a joint frame of reference. Thus, integration is achieved by a sequence of actions in the development and use of information across the supply chain.

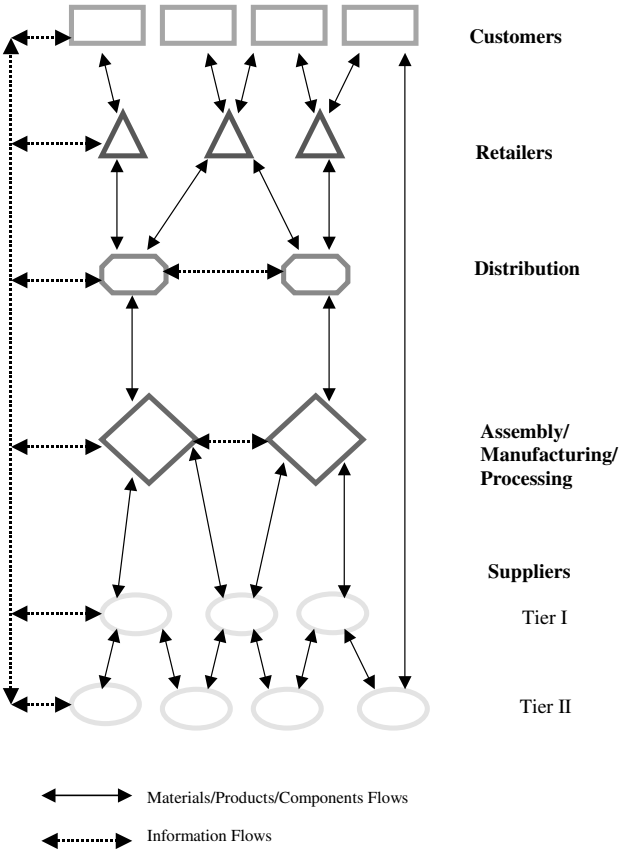


Figure 1: Integrated supply chain for an extended real-time enterprise.

One main approach to integration identifies four stages: information integration, planning synchronization, workflow coordination, and the recognition of new business models (Lee & Whang, 2001a). The first stage, information integration, requires the collection and sharing of a broad range of data regarding the status of operations in the supply chain. This includes data on sales, inventory, production, promotion plans, demand forecasts, and information about the location and delivery schedules of goods in transit. The second stage, planning synchronization, involves identifying how the newly developed and shared information is to be used. This requires not only the joint creation of decision-making standards, including common rules for defining outcomes and making choices, but also the application of this system to production planning, forecasting, replenishment, design, capacity utilization, and service. The third stage, workflow coordination, extends the activities in planning synchronization to the actual coordination of the operation of the supply chain, including production decisions, procurement and ordering, and product development and design. The final stage, recognition of new business models, takes advantage of the opportunities of the creation of an extended, real-time enterprise. This requires bringing customers into the information system and using customer information and involvement for real-time product customization, demand management, dynamic pricing, real-time quality control of manufacturing based on user experiences, reallocating capacity, developing new categories of information available from integrating the extended enterprise, developing new forms of service, and redefining the relationship between products and services.

Table 1 Efficiency Benefits from an Internet-Enabled Supply Chain

Make better forecasts
Reduce inventory risks and costs
Reduce procurement costs
Coordinate production, distribution, and fulfillment more effectively
More easily locate specific items in the supply chain
Monitor and respond more quickly to bottlenecks and other problems in the supply chain
Reduce lead time
Reduce delays and time lags in movement of components through the supply chain
Improve service while lowering costs
More rapid development of products
Faster time-to-market
Moderate bullwhip effect

Table 1 lists some of the most important gains we can expect from the new system of information integration across the supply chain. Perhaps the most significant is the ability to manage and control inventory levels, primarily by moderating the bullwhip effect. This is the term used to describe the tendency for small variations in demand by downstream end customers to result in increasing inventory variation across the various upstream stages of the supply chain (Simchi-Levy, Kaminsky, & Simchi-Levy, 2000).

The bullwhip effect is a consequence of decision-making using incomplete information, which comes from the absence of supply chain integration (Lee, et al., 1997). Typically, production decisions are made using information only from the next level in the supply chain, and this results in the use of hedging decision models to adjust for the uncertainty in this incomplete information. In other words, when individual actors in a supply chain use information known only to them (usually orders from the actors at the next level of the chain) to make forecasts and orders, and then pass only their orders on to the next actor in the chain, the result is that increasing levels of inventory are held at succeeding lower levels of the chain. When this incomplete information is combined with production lead times, use of large batch orders, and use of higher order size to protect against shortages, the end result is considerable overshooting or undershooting of optimum inventory levels. Thus, much of the bullwhip effect and its negative effects on inventory levels is the result of each level of the supply chain making demand forecasts and optimization decisions based on information affecting only a limited part of the entire chain (Simchi-Levy, 2000). When each unit makes independent production decisions based on independently generated forecasts, using incomplete data, the resulting inventory levels for each organization in the supply chain and cumulatively across the supply chain are excessive and inefficient.

By contrast, when each part of the supply chain obtains real-time information about actual end demand, and when inventory management decisions are coordinated, inventory levels are reduced across the supply chain. The key is to provide each unit of the supply chain with more complete information shared by all units. Exactly what kind of information is this? Demand forecasts, point of sale, capacity, production plans, promotion plans, and customer forecasts are some of the many forms this information can take (Lee & Whang, 1998). This more complete information helps reduce distortions and errors and permits better decisions. An important example is shared information about point of sale demand and demand forecasts, which permits development of a single demand forecast, or at least a coordinated series of forecasts based on common data, everywhere in the supply chain.

Different forms of cooperation are created through linking together the information systems of firms using the Internet. For example, we can distinguish between automatic replenishment programs (ARP), where inventory restocking of individual items is triggered by actual needs, and collaborative planning/forecasting/replenishment (CPFR), where firms engage in joint planning to make long-term forecasts that are updated in real time based on actual demand and market changes. CPFR involves a higher level of cooperation and collaboration than ARP does. Efficiency gains in CPFR come from reductions in overall inventory while increasing stock availability (especially during promotions) and achieving better asset utilization (Stank, Daugherty, & Autry, 1999). One reported success story is a CPFR trial at Nabisco and the grocery chain Wegmans Food Markets that resulted in increasing sales while significantly reducing inventory and improving service levels (Oliver, Chung, & Samanich, 2001).

The efficiency gains from the Internet arise from its capacity to quickly transmit large amounts of complex information throughout the supply chain. However, there are important choices about the nature of the Internet links to suppliers. Generally, two broad e-procurement options exist: first, use of broadly based transactional exchanges to aggregate sellers (and sometimes buyers) so as to create a larger market with lower search costs and lower product costs; and second, the use of Internet-based links to create a relationship exchange among established suppliers so as to increase information flow and manage inventory. There are important distinctions between these options relevant to management decisions (Kaplan & Sawhney, 2000). The great benefit of the broad transactional exchange is the ability to consolidate markets, expand access to suppliers, lower search costs, and lower acquisition costs. The online auction is perhaps the best example of such a system, with many buyers and sellers operating much like the stock market. The reverse auction is a variant, in which a buyer places a request for bids for a product and receives competing bids from a collection of suppliers. By contrast, the private trading exchange (PTX) or private hub is a term used to describe a relationship exchange used for exchanging information and automating transactions with a long-term supplier (Dooley, 2002; Gurbaxani, 2002).

The use of a broad transactional exchange for e-procurement has many potential benefits, especially if the product has commodity-like characteristics with many suppliers, many of which have additional capacity (Emiliani, 2000). Where price is the primary consideration, transactional exchanges can generate significant savings. Goods such as those for maintenance, repair, and operations (MRO) meet these criteria (Croom, 2000). However, there are important qualifications to the hoped-for benefits. Many of the public exchanges established to provide such markets failed to generate adequate aggregation of buyers or sellers (*The Economist*, 2001). Additionally, the realized gains from reverse auctions may be much less than expected. Firms using an online reverse auction may achieve significant gross savings, only to find those savings reduced by hidden costs associated with switching suppliers (Emiliani & Stec, 2002). The relationship with suppliers in an arm's length public exchange is generally adversarial, and this is compounded by the fact that use of an exchange for e-procurement sets suppliers against each other in a bidding war. Consequently, many suppliers have avoided these exchanges, thereby holding down liquidity. Attempting to fix these weaknesses, many arm's length transactional exchanges have developed new capabilities for adding value. This includes Free Markets' ability to provide specialized information needed for complex transactions, specialized solution providers like Biztro.com, and sell-side asset exchanges such as transportal network (Wise & Morrison, 2000).

Perhaps more significant, arm's length trading in an exchange is inconsistent with achieving greater collaboration with suppliers. In a mutually beneficial relationship with a supplier, negotiations over price, quantity, replenishment, product development, and variations in product can be used to establish parameters for automating actions over the Internet. The relationship exchange provides benefits beyond price. Savings can be realized through reductions in ordering costs, time saving, lower search costs, and procedures for facilitating the ordering process. These savings can be considerable, such as the reduction in transaction costs by British Telecom from $113 to $8 (Lucking-Reiley & Spulber, 2001). A relationship exchange can also be used to combine a transaction environment for suppliers with a complex interfirm communication system. Exostar is a defense aviation exchange supported by Rolls-Royce, BAe Systems, Boeing, Lockheed Martin, and Raytheon. This exchange is used to coordinate the relationship of suppliers and product design for complex systems, such as fighter aircraft (*Economist Intelligence Unit*, 2002).

The decision to use a B2B exchange (whether a transactional or relational) must link Internet strategy with the nature of the relationships that exist in the supply chain (Jap & Mohr, 2002). The obvious difficulties come when firms previously engaged in an arm's length transaction attempt to shift to a more collaborative environment or when firms in a collaborative environment are moved to a public exchange with more intense conflict.

MANAGEMENT ISSUES IN SUPPLY CHAIN COLLABORATION

The Internet does not create the need for cooperation and collaboration in supply chains; this has always been important. However, the Internet does provide new

opportunities to be won from collaboration and offers some potential ways to overcome barriers to collaboration. Most of the improvements in efficiency from the Internet happen only when the separate parts of a supply chain work closely together. There are considerable gains from collaboration, but there are perhaps as many hurdles to be overcome. Managers of supply chains must be even more aware of the barriers to collaboration and work to use the Internet to overcome these barriers. The demands of information sharing and collaboration in an extended, real-time enterprise are much greater than in traditional supply chains and will require organizational innovations.

Collaboration is difficult to achieve because different firms have different economic interests, goals, and ways of conducting business. Additionally, there are technical barriers, including problems in developing systems based on Internet standards that link legacy systems and provide a communication path between firms in the supply chain; integrating EDI systems to Internet systems; defining standards for content management; establishing XML and ebXML standards; providing common systems for logistics and procurement; and defining datamining standards.

Competing efforts by different firms to control the information system and capture profits from its use create conflicts among the members of a supply chain. The benefits of collaboration and coordination may flow disproportionately to one part of the supply chain, although these benefits can only be achieved through working together. Moreover, different firms have different levels of risk associated with their position in the supply chain, and this invariably affects the measures they use to define the operation of the supply chain and choices on production, resupply, promotions, and product development and innovation. The willing transfer of vital information by parties that could use it to achieve market advantage or improve the strategic position of a supply chain partner is problematic. An end seller will need to devote considerable resources to a complex information gathering system in order to obtain real-time point of sale data, and will share this information only for a fee. Achieving genuine information visibility across several organizations, given the variety of these constraints, is perhaps unattainable.

Simulations of supply chain operations help to identify many of the benefits and barriers to information sharing and coordination (Zhao, Xie, & Zhang, 2002). Several factors can influence the magnitude and distribution of the benefits, including different patterns of demand variation at the retail level (amount of demand fluctuation and whether demand is rising or falling through the period), different levels of capacity tightness (high and low), and different kinds of coordination and information sharing (no sharing, sharing of demand forecasts and order information, and amount of advance time for coordination of orders). Also the benefits themselves can be differentiated by costs for retailers, costs for suppliers, costs for the entire supply chain, service levels for suppliers, and service levels for retailers. The results of repeated simulations suggest a complex pattern of benefits. Information sharing is uniformly beneficial to all parties in the supply chain, with the best performance coming when order information is shared. Coordination of ordering benefits all parties but provides the greatest benefits to suppliers, who are able to improve capacity utilization with more order information and a longer planning horizon. Likewise, information sharing and order coordination under different demand patterns generate much larger gains for suppliers than for retailers, because suppliers are able to use the information to adjust capacity utilization. Although suppliers almost always experienced large gains, when demand is falling and capacity utilization low, retailers find their total costs rise and service declines under information sharing and order coordination. These differential benefits suggest some of the barriers to coordination and information sharing in a supply chain. Despite the large gains to the entire chain, the uneven distribution of the gains means some mechanism for redistribution of the gains is needed to reap these benefits.

Another useful way to highlight some of the barriers to collaboration is to examine the metrics needed to evaluate an Internet-enabled supply chain, and the issues related to reaching agreement about the metrics and trigger points for decisions. Efforts at collaboration engage the different needs, expectations, and interests of different firms in a supply chain. One firm might prefer an exclusive focus on measures such as its inventory turns or capacity utilization, whereas a much better approach for enhancing collaboration would be metrics that provide measures of performance across the chain and identify points where weakness can damage the entire operation. Thus, effective measures need to be multifunctional and cross enterprise. With an Internet-enabled supply chain, these kinds of measures are more easily attained. However, this means shifting the emphasis from the individual enterprise to the supply chain (extended, real-time enterprise) as a whole in its capacity to provide customer satisfaction (Hausmen, 2000). Getting separate firms to accept this kind of interchain thinking can be difficult; these difficulties become obvious when we consider some of the actual metrics.

The Internet creates new opportunities to interact with customers and to provide products more closely customized to individual customer preferences. However, supply chains may be organized so that only the firm facing the customer is focused on this opportunity. In this setting, customer service measures will be based on either "build-to-stock" or "build-to-order" products; different firms in a supply chain with different emphases will find agreement on such measures difficult. The firm building to customer order will want measures of rapid customer response time throughout the chain, but may need to deal with companies having a build-to-stock approach. Similarly, an integrated supply chain will need to provide inventory measures across the entire chain, not just at a single firm, and then use these measures to compete against other supply chains. Such measures must aggregate the inventory of upstream suppliers and downstream end customers. Measures of the speed of flows must also consider the entire system and each of its links to materials.

Obviously, developing this information, sharing it, and agreeing on its meaning for managing the supply chain can encounter many difficulties, primarily because firms are not in the same situation. A firm that centers its value proposition on low cost will need measures different from those of a firm that emphasizes customized products. Just

because the value proposition for the firm nearest the end customer uses an approach for measuring value, this doesn't mean all firms in the supply chain can or should use it.

Are there solutions to these problems? The Internet itself may present some new opportunities. The availability of real-time information providing visibility across the supply chain permits formulation of a wide variety of refined measures of supply chain operation and of the benefits of various choices. This high-quality information can be used to reach agreement where past conflicts were the result of poor information. Another possible solution to intrachain conflict is to recognize that in addition to the different economic interests of firms in a supply chain, there are also significant differences in economic power (Cox, Sanderson, & Watson, 2001). Buyers typically have more power than suppliers and may be in a position to define the terms of the supply chain and terms for information sharing and collaboration. Sometimes, as in the role of Intel or Microsoft as suppliers to the personal computer industry, this relationship can be reversed. This power can mean that buyers (or suppliers) may be able to use their strategic position to determine the division of profits from the information supply chain and extract the greatest profits.

Another possibility is to limit the type and scope of information shared and accept the now limited gains this can bring (Lamming, Caldwell, & Harrison, 2001). This way steps back from the ideal of complete supply chain integration, and is more likely to work in situations involving a long-term relationship where a series of negotiated arrangements through time that reflect the relative distribution of power in the chain and adjust to reflect the different economic interests of the parties emerges. The nature of this negotiated outcome may be some combination of an adversarial and collaborative relationship. One "federated" approach (Oliver et al., 2001) to the process advocates supply chain partners working to align their business objectives, "performance levels, incentives, rules, and boundaries" in conjunction with developing an understanding of the trade-offs relating to cost and service. This approach calls for an ongoing set of negotiations on broad objectives rather than a detailed one-time specification of measures and outcomes. The continuous flow of high-quality information facilitated by the Internet can help firms engage in this continuous negotiation over actions and needs.

This federated, or decentralized, approach is suggestive of the organizational innovations needed to realize the benefits of an Internet-enabled supply chain. The exchange of complex information in real time facilitates precisely this kind of continuous negotiation and adjustment of goals, operations, and measures. The main purpose of cooperation and collaboration is to reap the efficiency gains and the production flexibility inherent in adopting an Internet-based system. The specific meaning of these benefits is likely to be highly differentiated from one supply chain to another and even for different product variations in the same supply chain. Continuous negotiation can make these adjustments and help sustain collaboration. Further, the distributed decision-making environment of an extended, real-time firm may be much more conducive to collaboration than the vertical, centralized model (Tapscott, 2001).

NEW BUSINESS MODELS

Assuming the roadblocks to collaboration presented by technical problems and the differences in economic interests can be overcome, gaining the benefits of collaboration facilitated by the Internet also requires the development of new business models and practices. The application of Internet technologies expands the business options for the firm and the role of the supply chain in achieving its goals. For purposes of illustration, we will emphasize three:

1. The simultaneous development of products, processes, and supply chain design. This requires high-level collaboration but enables the various firms in the supply chain simultaneously to develop new products, configure the production processes, and design the supply chain. When a new product requires both a new set of production processes and changes in the configuration of the supply chain, simultaneous development permits these decisions to be coordinated, thereby reducing conflicting needs (Fine, 1998).
2. Increased ability to use and extract information from the supply chain, which makes possible efforts to integrate demand management and the supply chain, and the integration of supply chain management and customer relationship management (Lee, 2001).
3. The development of pull–push strategies to replace push strategies. This business model allows for reshaping the make/buy decision from the perspective of the extended enterprise and permits improved flexibility in developing the capacity for customization (Simchi-Levy, 2000).

The development of an extended, real-time enterprise through the Internet requires the supply chain be reconceptualized into three chains—organizations, technologies, and capabilities—and used to rework the way products are developed in relation to the production processes across the supply chain (Fine, 1998). This involves understanding the assets that exist across the supply chain: knowledge assets, integration assets, decision system assets, communication assets, information extraction and manipulation assets, and logistics assets. A key element of an extended enterprise strategy is to direct the design and redesign of the supply chain in relation to existing and new products so as to create competitive advantages. Resources for competitive advantage reside across the entire supply chain, and the chain itself must be designed so as to take advantage of these assets. Through the Internet, rich information can be exchanged among various tiers of the supply chain, which permits the identification of emerging assets and their organization into increasing competitive advantage. The most important result of this business model is the ability concurrently to create products, develop production processes, and design supply chains. This capability permits the solution of potential manufacturing problems in the supply chain at the point of its design and in conjunction with design of the product and thereby reduces time to market. In a competitive world

with short product cycles and a high premium on flexibility and customization, this capability provides enormous competitive advantage.

The ability to integrate product development, process, and supply chain design alters the choices to either make or buy an element of the product (Fine, Vardan, Pethick, & El-Hout 2002). As always, firms must decide where the greatest value is generated in the supply chain and distinguish between areas of internal competence and external dependence. The Internet can affect these decisions. Because the value of rich information available in real time rises so dramatically, even as its cost falls dramatically, control over the creation, distribution, and use of information in the supply chain may be where the greatest value lies. Traditionally, decisions about make or buy were defined in terms of the strategic value of the component or process, as in the value added, the importance to the customer, or the knowledge value. In an e-business environment, where rich information is plentiful and easily shared, and where products, processes, and supply chains are codeveloped, the core competency may reside in the ability to orchestrate the organizations, technologies, and capabilities in a supply chain. The ability to anticipate market opportunities and configure or reconfigure a supply chain to respond rapidly and flexibly may replace traditional make/buy decisions. This may mean that what remains consistently in-house are the knowledge assets associated with technology, markets, and supply chain design.

A second illustration as to how an Internet-based supply chain can have a significant impact on strategy and new business model development is in the management of demand (Lee, 2001). The ability to link real-time demand information through the supply chain gives rise to the capacity to manage the relationship between customer needs and supply chain capabilities, so as to increase benefits for both. The key is to recognize actions taken by the firm to influence demand, and to integrate those actions with the capabilities of the supply chain. In the most elementary sense, this means that marketing efforts to boost demand should consider the impact over time on the actual shape of the demand profile and whether this is consonant with an optimized supply chain. The cost of expanding capacity to meet induced demand spikes may exceed the revenues generated in extra sales. The solution is to have a clear and precise sense of supply chain costs associated with different levels of production, and integrate those costs with decisions relating to promotions. Further, promotions need to be coordinated across the supply chain, so that promotions at the retail level are linked to promotions at the wholesale level. Internet-based systems achieve this coordination for individual SKUs and at retail, wholesale, and production sites. The system can take into consideration the impact of decisions on one product as they ripple over other products, production, and logistics.

A final area for new business models that derive from Internet-based supply chains is the ability to substitute a push–pull strategy for a push strategy (Holweg & Pil, 2001; Simchi-Levy et al., 2000). In an era where information about final demand was not available across the supply chain, production decisions were based on demand forecasts supplemented by order information from the next level in the chain. Products of limited differentiation were manufactured based on aggregate long-term forecasts and pushed forward to consumers in the usually unrealized hope that the forecast was correct. By contrast, a pull strategy involves a direct sales and build-to-order capability in which customers order products configured to their specifications (that is the meaning of highly differentiated). This is a very unusual situation; a more realistic option enabled by Internet technology is to redefine the boundary between push and pull. Typically, some production must take place prior to the receipt of actual orders because of lead times and the fact that customers typically want quick delivery. This is true even in a direct sales system with end demand shared across the supply chain. Thus, one part of the supply chain will need to be organized around the push model while the other part will be organized around the pull model. The goal of many firms is to move beyond complete reliance on a push strategy, and define the most efficient point in the supply chain to locate the boundary between production to stock (push) and production to order (pull).

The availability of demand information across the supply chain expands the options for locating this point. Availability of final demand information in real time makes it possible for firms throughout the chain to generate better forecasts and reduce demand uncertainty. Forecasts are still necessary at the push level of the production process, but real-time demand information permits more accurate and more differentiated forecasts. This is because long-term forecasts can be updated and adjusted based on more accurate information, rather than the bullwhip affected information found in the traditional supply chain. Further, firms have long relied on aggregated forecasts because they were more accurate than forecasts of specific products. However, this forces greater reliance on push strategies, with the attendant costs from unsold goods and lack of flexibility.

Use of real-time information about final demand permits more differentiated forecasts, because final demand information is about specific products, and this information can be use to spread pull strategies further down in the supply chain. Differentiated forecasts permit push production strategies using specific product information, and permit the development of pull strategies at points farther upstream in the supply chain. More generally, real-time information about final demand permits the design of the supply chain so as to integrate and optimize considerations associated with both push and pull. Push focuses on production scale, distribution logistics and timing, lead times, inventory, and transportation. Pull focuses on flexibility, customization, service levels, and delivery. The trade-offs between push and pull still exist, but are shifted in such a way that both can be achieved in more effective ways (Holweg & Pil, 2001).

As we have seen, one important goal in adopting an Internet-based supply chain is to shift the push–pull boundary so as to provide a more responsive and even agile supply chain. This is especially important when demand is uncertain as to volume and variety of product and when the supply base is affected by uncertainties related to process and technology of production (Lee, 2002). One

company facing such a situation is Xilinx, which designs high-end and customized semiconductors. The market for application-specific semiconductors is highly variable and the production process is technologically sophisticated and difficult. Xilinx operates without fabrication facilities, using instead close partnerships with fabrication foundries in Asia. The production process is split between a push level and two different opportunities for production based on pull. First, the fabrication of the chip is separated between an initial phase and a final phase, which are carried out by different foundries. The final configuration of the chip is delayed until actual demand is known, which permits Xilinx to respond to shifts in customer needs. Second the Internet is used for customized configuration of products even after customers have received shipment (Lee & Whang, 2001b). The field-programmable logic devices made by Xilinx, and used in products like communication satellites, need continual changes in configuration. In normal circumstances this would require frequent on-site service and replacement. However, Xilinx has developed the ability to use the Internet to modify and upgrade these devices after they have been delivered to customers, in effect extending the pull model to an after delivery capability.

Another example of a firm making extensive use of Internet-based communication and collaboration with its suppliers is Adaptec (Hauseman, 2000; Lee & Whang, 2001b). The "fabless" semiconductor manufacturing business model requires close connections with suppliers, a significant challenge from both a cost and performance perspective, especially given that the elements of the supply chain are dispersed around the globe. Adaptec uses specialized communication software for the Internet to link its customers to the design, production, assembly, and packaging stages of the supply chain. The information includes purchase orders, production forecasts, shipment schedules, prototype specifications, and test results. In a market based on rapid response to custom demand, Adaptec has used this information system to cut in half its cycle time for new product development. The relationship with suppliers has helped to generate trust, which facilitates continuing investment in the technology needed to make the information system work.

CASE STUDIES

Nortel

A number of firms have adopted some or all of the operations of e-business in an extended, real-time enterprise. Nortel is a good example of a firm that rearranged its business model and supply chain using the Internet in order to cope with rapidly changing markets (Fischer, 2001). The anticipated growth of fiber-optic networks offered an opportunity for Nortel, but its realization required much more agility in its production and supply chain system. Nortel sold off production facilities and partnered with its component manufacturers. Because the product required extensive customization, each customer is now assigned to a dedicated, but virtual, supply chain. This means that Nortel's role is to work directly with each customer to design and configure products, and communicate this information in real time to the Nortel supply chain partners. The people at Nortel who design systems also communicate with suppliers, an example of linking product design and the supply chain by customizing each simultaneously. Design of products also includes suppliers in early stages, made possible by the increased communication flows. The benefits include shorter time for suppliers to develop bids, greater willingness by suppliers to commit capacity to Nortel because of visibility into Nortel customers' end demand and Nortel's strategy, better understanding of supply chain capabilities by Nortel and its customers, and more informed decisions about trade-offs, shifting the push–pull boundary to permit delayed configuration of products, and more rapid response to orders. At the same time, however, inventory has not declined, and Nortel is still not as agile as some of its competitors in the fiber-optic market.

Two other firms that have come to define many of the benchmarks for implementing an Internet-based integrated supply chain strategy are Dell Computer and Cisco Systems. These firms provide blueprints for how integrated systems can be constructed, along with the benefits and the perils of such systems. Each firm is an example of a radical outsourcing strategy based on the construction and integration of a global supply chain.

Cisco Systems

Cisco's products are quite complex and include the hardware (large-scale routers, LAN switches, and WAN switches), software, service, and integration capabilities to implement an end-to-end network solution for an enterprise. The extraordinary growth and the rapid technological changes of the networking market present significant challenges for Cisco. In response, it has adopted an aggressive strategy for developing new technology and knowledge assets. In addition, Cisco created a remarkable networking system to link its customers, employees, and the myriad collection of assemblers, suppliers, semiconductor manufacturers, logistics, and service providers who make up the Cisco system. Cisco's acquisitions and the creation of an extended, real-time enterprise have made possible rapid expansion to meet sales growth and technology changes.

A central feature of Cisco's overall business strategy was the creation of an integrated information system from customer to supplier, and selection of strategic partners and suppliers to populate its supply chain. Indeed, Cisco is one of the best examples of the use of the Internet to create a collaborative system designed to benefit all members of the system. The main goal is the performance of the supply chain, not merely obtaining the lowest price from suppliers. Internet-defined protocols are used for all communications across the system and common applications throughout the company. This permits high levels of interoperability of communication capabilities, information systems, decision support systems, collaboration systems, employee access to information, and rapid access to all information concerning a customer.

The implementation of this system at Cisco began in 1993, and the most recent version is the e-Hub (Grosvenor & Austin, 2001). This is a private e-marketplace that

Table 2 Benefits of Cisco's Extended Supply Chain Information System

90% of business transacted over the Internet
Reductions of 45% in inventory as a percent of sales
Order cycle time reductions of 70%
Time to volume—the time required by production lines to scale for mass manufacturing—cut by 25%
Improved service system using Internet-based self-service
70% of customer service inquiries resolved online
Higher customer satisfaction levels
Rapid creation and marketing of new products
Ship directly from the manufacturer
Ability to scale up continuously to meet rapid market growth
Ability to reallocate employees to higher productivity jobs (double the revenue per employee as competitors)

facilitates information flows across the entire supply chain. The e-Hub permits information sharing about demand forecasts, supply status updates, event alerts, and inventory levels for components, and provides the basis for collaborative planning and execution. Additional features include an Internet-based customer service and support system, product testing using the Internet, the use of "dynamic replenishment" software to link customer orders to supply chain producers, and a sophisticated system for data mining of the information system. The benefits of this system are considerable and are outlined in Table 2. By early 2000, Cisco had the largest market capitalization of any firm in the world.

In the wake of a steep decline in spending for networks, however, this remarkable system floundered and then broke down The capacity of the Cisco supply chain to scale up to meet market growth did not work well in reverse; that is, Cisco was unable to scale down as quickly or effectively. The result was an enormous buildup of inventory and a $2.5 billion inventory write down against Cisco's earnings.

The problem, according to one study (Lakenan, Boyd, & Frey, 2001), came from the misalignment of goals and business plans between Cisco and its supply chain partners, who are mostly contract equipment manufacturers. Various units of the supply chain used different standards for making decisions, and the informal communication that could have uncovered these differences did not happen. The formal contractual relationships between Cisco and its suppliers were unable to capture the tacit rules necessary for an effective relationship, and the information flowing through the supply chain was not rich enough to make adjustments to changing market conditions. The source of the differences was a mismatch between Cisco's need for production flexibility and its suppliers' need for predictability, which made it difficult to turn off the push part of the production system fast enough to adjust to deteriorating market conditions. What could have been done differently? According to Lakenan et al., firms like Cisco need to think in terms of a supply web with differentiated production capabilities and greater flexibility of capacity utilization. This involves not only more flexibility in product design but also a greater willingness for strategic supply partners to make adjustments and adaptations to the supply web as conditions warrant.

Dell Computer

Dell is rightfully seen as one of the closest approximations of an extended, real-time enterprise. Dell's business model has always been direct sales to end users, who are primarily enterprises (Mendelson, 2000). Its role is to assemble and rapidly deliver the final product based on customer configuration, with all components outsourced. A large and growing portion of Dell's sales are made over the Internet, which also is the medium for much of Dell's customer support and a key element of the integrated information system linking customers to the supply chain. Except for a brief interlude, Dell has always been a direct sales firm that provided build-to-order machines, which facilitated its adoption of the Internet for sales. Additionally, the initial emphasis on speed of operation and low inventory also contributed to the adoption of the Internet. The direct sales model already contributed valuable information about end demand and market shifts, along with indicators about how service and support could be used to add value to the product. In many ways, Dell's operations were defined in Internet terms before the emergence of the browser-based Web. Dell was already an information-intensive firm that used this information to accelerate the speed and leanness of its operation.

The Internet built on and accentuated Dell's business model and corporate culture. This helps account for Dell's status as the primary example of an Internet-based firm. Dell's position rests on its ability to create and manage an information system that integrated the production and delivery of components, assembly of machines, and direct sales, distribution, and servicing of those machines. One measure of the ease of transition to the Internet was the rapid jump in Internet sales from $1 million per day in early 1997 to $50 million per day in 2000 (Kraemer & Dedrick, 2001). At the same time, the transition brought significant increases in efficiency and customer satisfaction.

Much like Cisco, Dell's Internet capabilities extend from customer to suppliers. The ability to custom configure the machine effectively brings the customer inside Dell. Further, the Internet permitted Dell to provide customers with rich information about the product, including the cost for each variation of the machine. Customers became much better informed about the product and the purchase experience, and felt like they were on the shop floor making choices and understanding those choices. The Internet provided better tracking information to customers at a lower cost to Dell than the telephone. Also service was made easier by using the Internet to provide downloads and even to diagnose problems. The development of a customized Web site for customers provided Dell with asset management capabilities. The Internet-based volume of rich information about customers allowed Dell to seek out and use this customer information to develop products and services. As a result, marketing strategies can be quite differentiated and the Internet can

be used to get instant feedback on the effectiveness of these strategies.

The Internet-based information system extends to the supply chain base in much the same way as it extends out to customers (Kraemer & Dedrick, 2001). The Dell supply chain is composed of a relatively small number of suppliers, each with a very strong relationship with Dell. The supply chain is global, as is the assembly system. Dell receives real-time information from its suppliers about their capacities, inventory, quality measures, and costs, and suppliers receive information from Dell on demand forecasts, sales, quality measures, and customer needs. The assembly plants receive orders every two hours and establish a new production schedule. Suppliers maintain component inventories close to the assembly plants. Typically, Dell is able to assemble an order in one day (Holweg & Pil, 2001).

Dell ties this supply chain together with the information that comes from customers. First-tier suppliers are able to see Dell's demand forecasts (updated through an Internet portal) and are able to gauge their operations accordingly. Providing direct information about sales and demand forecasts in real time to suppliers mitigates the bullwhip effect. Suppliers, aided by real-time information, are in a position to produce for end demand and make adjustments separated only by lead time (Kraemer & Dedrick, 2001; Kraemer, Dedrick, & Yamashiro, 2000). The information system also permits some suppliers to ship parts of the end product directly to the customer, thus avoiding shipment to the Dell plant and reshipment to the customer.

For Dell itself, all inventory is work in progress; there is no finished goods inventory, since products are shipped as soon as they are assembled and tested. This Internet-based information integration has significantly contributed to the remarkable decline in inventory, from an average of about 35 days of component inventory on hand in 1995 to less than 4 days in 2001 (Kraemer & Dedrick, 2001). Dell also provides information about product defects for use at the point of production to improve quality control. Integration of the supply chain is enhanced by suppliers who are in close physical proximity to Dell, holding inventory within a few minutes of the assembly factory. Information is also used to establish very tight integration between delivery and service. The integrated supply chain permits a very simple yet powerful push–pull system. Information about end demand is used to pull the individual product through the supply chain system and to push components to the assembly floor in a flexible and adjustable manner. However, the Dell system does not reach the high standard of a build-to-order system, primarily because suppliers are required to maintain inventory levels and do not have complete access to customer orders and therefore cannot base their production on this information (Holweg & Pil, 2001).

Dell exemplifies a systemic understanding of how supply chain management must operate in an extended real-time firm. The benefits include reduced costs for sales, service, and supply chain operations, plus improved quality, delivery time, and service satisfaction. Dell's performance is better than that of its less Web-enabled competitors, and its broad capabilities have been essential in increasing its market share. Nonetheless, Dell has also experienced some of the same inventory problems as Cisco as a result of a misalignment with suppliers.

NMS Communications

The experience of NMS Communications helps to illustrate many of the issues in developing an Internet-based supply chain strategy. In particular, this case depicts the consequences of integrating the supply chain with near real-time information and the impact of this information on the operation of the firm. NMS produces printed circuit board assemblies for the telecommunications industry to enable transmission of voice over Internet protocol (Arntzen & Schumay, 2002). Manufacturing of its products and new product prototypes is completely outsourced. Similar to Cisco and Dell, this is a business with high rates of technological change and the need for rapid delivery and quick time-to-market for new products. However, NMS operated its supply chain based on a traditional build-to-forecast system. This led to many expectable inventory and product development problems: excess inventory, parts shortages, long customer order cycle times, low delivery predictability, high end-of-quarter peak loads, and slow new product development. Recognizing many of these problems, NMS developed a plan to shift to an Internet-based, build-to-order supply chain and reposition itself as the integrator of the entire supply chain. NMS wanted to become the core of an extended, real-time enterprise. The effort to achieve this led to an understanding of the enormous changes in internal thinking and external relationships required by this plan.

To support these changes, NMS created a "hub-and-portal" IT architecture to integrate its internal systems and connect electronically with key customers and suppliers. The portal is multimode (it operates using XML, EDI, Web-based, and extranet capabilities) and supports the integrations of business processes, along with visibility and collaboration across the supply chain. Customer orders are posted to a Web site available to manufacturers and suppliers. The consequences for NMS and its relationship to customers and suppliers were substantial. These are summarized in Table 3.

NMS found the transition process to be difficult, primarily because of the scope of the changes required. Clearly, operating in near real time as an Internet-based, build-to-order firm requires rethinking much of operations. At the same time, many benefits and competitive advantages are also available.

FUTURE TRENDS

What is the future role of the Internet in supply chain management? Adoption of Internet-based business-to-business systems has been slower than many predicted, perhaps affected by the bursting of the dotcom bubble. What trends can we expect? The projection of one study (Carter, Carter, Monczka, Slaight, & Swan, 2000) is that before 2010 the Internet and World Wide Web will become the "backbone of electronic purchasing," integrating supply chains by providing the medium for all basic and essential transactions and communications, operating as the basis for access to "critical" information and

Table 3 Consequences to NMS Communication of Real-Time Operation

Changes in the process of production
Reduction of inventory levels meant that much greater agility was needed to deal with variations in demand and the uncertainties of the supply chain
Increased agility was accomplished through the following:
1. Greater modularity in components.
2. The time required to shift from producing one product to another was reduced from six hours to a few minutes.
3. The ability to see through to actual demand gave suppliers and manufacturers the ability to adjust production based on real-time data and minimize the use of forecasts.
4. Production shifted from a two-week schedule based on forecasts to a daily schedule based on actual orders.
5. The push–pull boundary was altered significantly. Where once customers pulled products only from finished goods and the forecast governed a large-batch manufacturing system, now customers would pull out of the vendor's raw materials inventories (still partly based on forecasts) into a flexible manufacturing system.
Changes in customer relations
NMS and the contract manufacturer can now use the information about orders and production capabilities and work with customers placing large orders.
Similarly, real-time communication across the supply chain permitted an iterative process of demand management between the NMS sales department and the contract manufacturer.
Demand management strategies based on lead times for production were implemented to reduce the variability of demand.
Virtually all customers were eager to participate in the new system.
Changes in product development
Delays and flaws in new product development not discovered until production had begun. This was caused by poor communication between product design at NMS and the production engineers at the contract manufacturer. The implementation of the Internet-based communication system led to a reduction of six weeks in new product development time.

measures of performance, facilitating complex partnering relationships that will provide the core competencies of many firms, supporting an expanded role for "demand-pull" business models, increasing the global reach for supply chains, and altering the skill set required of managers of supply chains.

Achievement of these predictions will require solutions to a multitude of problems. The integration of information systems is a considerable technical challenge; barriers to cooperation are created by conflicting interests among supply chain partners; goals are misaligned due to poor communication, even though the environment permits rich information flows; and management of a system with constant iterations and improvisations is complex and difficult, especially across multiple enterprises. Nonetheless, there are compelling business reasons for adopting the model of an Internet-based business in supply chain management.

GLOSSARY

Bullwhip effect The tendency for small variations in demand by downstream end customers to result in demand amplification and increasingly larger inventory levels and inventory variation across upstream stages of the supply chain.

Demand-based management Actions taken by a firm to influence the overall level of demand and the mix of products purchased to account for the needs and capabilities of the firm that supplies those goods.

E-business The marketing, buying, selling, delivering, servicing, and paying for products across nonproprietary networks that link an enterprise with its prospects, customers, agents, suppliers, competitors, allies, and complementors.

E-procurement The use of the Internet to purchase supplies through auctions, trading networks, or collaboration with long-term suppliers.

Extended enterprise The operation of a business through use of an integrated information system that aligns the various firms to permit collaboration with customers, strategic partners, and other parties in the supply chain.

Push–pull strategies Different strategies of production: push is production based on a forecast of demand; pull refers to production to fulfill a specific order.

Real-time firm The operation of an extended enterprise so that information about orders, operations, suppliers, and logistics is available in real time and permits new forms of management and relationships with customers and suppliers.

Supply chain The system of all organizations and steps in production, assembly, exchange, information transactions, financial flows, physical movement, and coordination from raw materials to final product.

Supply chain collaboration The coordination of decisions in a situation of coupled outcomes to produce plans superior to those available without coordination of decisions.

Supply chain integration The coordination and synchronization of supply chain operations through the sharing of information in real time by all parties.

Supply chain management The organization, design, and optimization of the business processes and the

physical and information networks that link raw materials to the production and delivery of end products.

XML (extensible markup language)/ebXML A computer language with specialized tags that permit coding for context and meaning for the information in a Web page. The ebXML effort is designed to provide standards for e-business tags.

CROSS REFERENCES

See *Developing and Maintaining Supply Chain Relationships; Electronic Commerce and Electronic Business; Electronic Procurement; Extensible Markup Language (XML); International Supply Chain Management; Managing the Flow of Materials Across the Supply Chain; Strategic Alliances; Supply Chain Management; Supply Chain Management Technologies*.

REFERENCES

Arntzen, B. C., & Shumway, H. M. (2002, January–February). Driven by demand: A case study. *Supply Chain Management Review*, 34–41. Retrieved April 24, 2003, from http://www.manufacturing.net/scm/index.asp?layout=articleWebzine&articleid=CA197691

Carter, P. L., Carter, J. R., Monczka, R. M., Slaight, T. S., & Swan, A. J. (2000). The future of purchasing and supply: A ten-year forecast. *Journal of Supply Chain Management, 36*(1), 14–26.

Cox, A., Sanderson, J., & Watson, G. (2001). Supply chains and power regimes: Toward an analytic framework for managing extended networks of buyer and supplier relationships. *Journal of Supply Chain Management, 37*(2), 28–35.

Croom, S. R. (2000). The impact of Web-based procurement on the management of operating resources supply. *Journal of Supply Chain Management, 36*(1), 4–13.

Dooley, P. (2002, May 15). Automated unattended b2b replenishment. *Ascet, 4.* Retrieved April 24, 2003, from http://www.ascet.com/documents.asp?d_ID=1099

Economist Intelligence Unit (2002, December 2). Europe: B2B revisited. Retrieved April 24, 2003, from http://www.ebusinessforum.com/index.asp?layout=rich_story&doc_id=6234&categoryid=&channelid=&search=B2B+revisited

Emiliani, M. L. (2000). Business-to-business online auctions: Key issues for purchasing process improvement. *Supply Chain Management: An International Journal, 5*(4), 176–183.

Emiliani, M. L., & Stec, D. J. (2002). Realizing savings from online reverse auctions. *Supply Chain Management: An International Journal, 7*(1), 12–23.

Fine, C. H. (1998). *Clockspeed*. Reading, MA: Perseus Books.

Fine, C. H., Vardan, R., Pethick, R., & El-Hout, J. (2002). Rapid response capability in value chain design. *Sloan Management Review*, 69–75.

Fischer, L. M. (2001). From vertical to virtual: How Nortel's supplier alliances extend the enterprise. *Strategy ± Business*, 1–8.

Grosvenor, F., & Austin, T. A. (2001, July/August). Cisco's eHub initiative. *Supply Chain Management Review*. Retrieved April 24, 2003, from http://www.manufacturing.net/scm/index.asp?layout=articleWebzine&articleid=CA154379

Gurbaxani, V. (2002, February). Partner integration networks: An economic analysis of the make vs. buy decision. Retrieved April 24, 2003, from http://supplychain.ittoolbox.com/documents/document.asp?i=1473

Hausmen, W. H. (2000, December). Supply chain performance metrics. Stanford Global Supply Chain Forum. Unpublished manuscript.

Holweg, M., & Pil, F. K. (2001) Successful build-to-order strategies: Start with the customer. *Sloan Management Review*, 74–83.

Jap, S. D., & Mohr, J. J. (2002). Leveraging Internet technologies in B2B relationships. *California Management Review, 44*(4), 24–38.

Kaplan, S., & Sawhney, M. (2000, May–June). E-hubs: The new B2B marketplaces. *Harvard Business Review*, 97–103.

Kraemer, K., & Dedrick, J. (2001). *Dell Computer: Using e-commerce to support the virtual company* (University of California—Irvine Case). Irvine, CA: University of California.

Kraemer, K. L., Dedrick, J., & Yamashiro, S. (2000). Refining and extending the business model with information technology: Dell Computer Corporation. *Information Society, 16,* 5–21.

Lakenan, B., Boyd, D., & Frey, E. (2001). Why Cisco fell: Outsourcing and its perils. *Strategy ± Business*, 54–65.

Lambert, D. M. (2001). Supply chain management: What does it involve? *Supply Chain and Logistics Journal, 4*(4). Retrieved April 24, 2003, from http://www.infochain.org/quarterly/F01/Lambert.html

Lamming, R. C., Caldwell, N. D., & Harrison, D. A. (2001). Transparency in supply relationships: Concept and practice. *Journal of Supply Chain Management, 37*(4), 4–10.

Lee, H. L. (2001, September). Ultimate enterprise value creation using demand-based management. *Stanford Global Supply Chain Management Forum*. Retrieved April 24, 2003, from http://www.stanford.edu/group/scforum/

Lee, H. L. (2002). Aligning supply chain strategies with product uncertainties. *California Management Review, 44*(4), 105–119.

Lee, H. L., & Whang, S. (1998). *Information sharing in a supply chain* (Research Papers Series, No. 1549). Stanford, CA: Stanford University Graduate School of Business.

Lee, H. L., & Whang, S. (2001a). E-business and supply chain integration. *Stanford Global Supply Chain Forum*. Retrieved April 24, 2003, from http://www.stanford.edu/group/scforum/

Lee, H. L., & Whang, S. (2001b). Winning the last mile of e-commerce. *Sloan Management Review, 42*(4), 54–62.

Lee, H., Padmanabhan, V., & Whang, S. (1997). The bullwhip effect in supply chains. *Sloan Management Review, 38*(3), 93–102.

Lucking-Reiley, D., & Spulber, D. F. (2001). Business-to-business electronic commerce. *Journal of Economic Perspectives, 15*(1), 55–68.

Mendelson, H. (2000). *Dell direct* (Stanford Business School Case). Stanford, CA: Stanford University Business School.

Nolan, P. (2001). *China and the global business revolution*. New York: Palgrave.

Oliver, K., Chung, A., & Samanich, N. (2001). Beyond utopia: The realist's guide to Internet-enabled supply chain management. *Strategy ± Business*, 1–10.

Siegele, L. (2002, February 2). How about now? A survey of the real-time economy. *The Economist*, 3–20.

Simchi-Levy, D. (2000, November–December). The master of design. *Supply Chain Management Review*, 74–80.

Simchi-Levy, D., Kaminsky, P., & Simchi-Levy, E. (2000). *Designing and managing the supply chain*. Boston: Irwin McGraw Hill.

Stank, T. P., Daugherty, P. J., & Autry, C. W. (1999). Collaborative planning: Supporting automatic replenishment programs. *Supply Chain Management: An International Journal, 4*(2), 75–81.

Tapscott, D. (2001, Third Quarter) Rethinking strategy in a networked world. *Strategy ± Business*, 34–41.

The Economist (2001, May 19). B2B exchanges: Time to rebuild, 55–56.

Weill, P., & Vitale, M. R. (2001). *Place to space: Migrating to ebusiness models*. Cambridge, MA: Harvard Business School Press.

Wise, R., & Morrison, D. (2000, November–December). Beyond the exchange: The future of B2B. *Harvard Business Review*, 86–96.

Woodall, P. (2000, September 21) The new economy. *The Economist,* 5–40.

Zhao, X., Xie, J., & Zhang, W. J. (2002). The impact of information sharing and ordering co-ordination on supply chain performance. *Supply Chain Management: An International Journal, 7*(1), 24–40.

Supply Chain Management Technologies

Mark Smith, *Purdue University*

INTRODUCTION

Supply chain management (SCM) attempts to identify the most cost-effective or profitable way of "getting the right product to the right place at the right time" (Bendiner, 1998). It is concerned with the goal of delivering the right amount of products or services from the point of origin to the point of consumption in the least amount of time and at the least cost. This becomes complex in supply chains that involve multiple levels of suppliers, manufacturers, distributors, retailers, and customers. A well-managed supply chain benefits all the members of the chain, maximizing throughput and profit, by effectively coordinating all of the resources in the chain.

Traditionally, manufacturers focused on the production operations of a business. The object was to get products through the manufacturing plant as efficiently as possible. Today this has changed. Large companies no longer control the entire supply chain. Businesses that used to do all their manufacturing operations internally have become finished goods assemblers, purchasing as much as 70% or more of their components from suppliers. Many components come from other countries, requiring the supply chain to accommodate an international scope. Trends toward globalization have introduced issues of exchange rates and tax considerations. The supply chain paradigm has shifted from manufacturing-centric management to customer-centric management and has been extended to cover many more participants in the chain. The information flow along the chain has become the dominant factor in implementing supply chain improvements. These changes and the extraordinary growth of the Internet have made the field of SCM ripe with opportunities for competitive advantage using information technology tools and packages.

By analyzing the principles of supply chain management, one quickly sees that information technology (IT) plays a strategic role both directly and indirectly in a successfully managed supply chain. Looking at the items in Table 1, one can make the following observations:

- To segment customers based on their needs, data describing their buying preferences, quality and cost concerns, geographic locations, and other demographic information must be analyzed. This requires data mining technologies from both internal and external database systems.
- Customizing the logistics network is done most effectively using supply chain simulation and optimization software tools.
- Signals from market demand are best captured and analyzed using demand management technologies.
- To differentiate the product closer to the point of actual consumption, the demand planning system must know in real-time what the customer wants, and advanced planning and scheduling technologies must extend the visibility of production schedules to the company's suppliers, so that components are there when needed.
- To source strategically, supply planning and execution systems that track and rank vendor performance in terms of cost, quality, and on-time delivery must be in place. Supply chain simulation software may also be used to make decisions regarding geographic locations of suppliers and raw materials.

Table 1 Seven Principles of Supply Chain Management

1. Segment customers based on customer needs.
2. Customize the logistics network.
3. Listen to signals of market demand and plan accordingly.
4. Differentiate product closer to the customer.
5. Source strategically.
6. Develop a supply chain wide technology strategy.
7. Adopt channel-spanning performance measures.

Source: Anderson, 1997

- The technology strategy for SCM is dependent on technology infrastructure decisions. These decisions must be made in light of the entire supply chain.
- Channel spanning performance measures can be successfully implemented and monitored only if the technology is in place to allow timely tracking and control.

WHAT IS THE SUPPLY CHAIN?

The supply chain for a product (or service) is the system of companies and business functions that it goes through, from creation to delivery to the ultimate customer. For a typical manufacturing company, the supply chain might be modeled as follows:

Suppliers ⇒ Manufacturer ⇒ Distributor ⇒ Consumer

Most companies, however, have much more complexity in their supply chains. This includes multiple levels of suppliers (the supplier's suppliers) and multiple levels of distribution and finished goods warehousing before the product gets to the ultimate customer (Figure 1).

As mentioned earlier, the traditional production management approach has been internally focused on the manufacturing process within the firm. These efforts focused on improving production efficiencies and scheduling. The goal of SCM is to eliminate costs from inventories and shorten delivery times by developing closer linkages between the production and actual consumption of the product. This linkage strategy is quickly becoming one of the primary competitive advantage strategies of business today.

Without an extended SCM focus, each company in the chain manages delivery service requirements by building inventories, without regard (information) to what others in the chain are doing. Unfortunately, this adds cost to the product for storage and obsolescence when the product is not consumed as rapidly as it is produced, or for expediting product when it is consumed more rapidly than

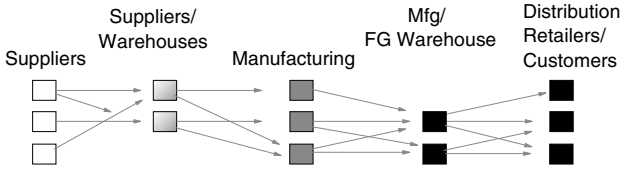


Figure 1: Extended supply chain model.

expected. This is true at each link within the supply chain. Each member attempts to optimize inventories based on demand history or a limited view along the chain, using inventory management techniques like economic order quantity. This can produce a phenomenon called the "bull-whip effect" (Lee, Padmanabhan, & Seungjin, 1997), which results in unexpected demand patterns for suppliers from changes further down the chain.

The shift to a customer-centric paradigm, which has occurred over the last several years, has changed the way the supply chain is managed. More focus on the customer has placed emphasis on ways to reduce product cost and at the same time shorten cycle times throughout the entire chain. Managing the entire supply chain as a single entity is the goal of current SCM techniques. The ability to do this is only available through expanded use of information technology.

The benefits of effective SCM can be great in terms of lower inventory costs, better production scheduling, and ultimately higher profits for the participants. To make it work effectively, members of the chain must have a great deal of trust in one another. This is not easy to accomplish, however, because the traditional attitude between businesses has been that negotiations must produce a winner and a loser (Poirier & Reiter, 1996). Additionally, the practice of "squeezing suppliers" for price reductions by dominant members of the supply chain has exacerbated the problem.

To better understand SCM technologies and how they can be used to overcome some of these problems, a better understanding of exactly what SCM is and how it functions is helpful. A quick review of Michael Porter's value chain analysis model (Porter, 1985) is a good place to begin this discussion.

SUPPLY CHAIN MANAGEMENT AND THE VALUE CHAIN

Porter's value chain model depicts activities within a firm or industry that show how value is produced through the creation of a product (or service). Primary activities include inbound logistics, operations, outbound logistics, marketing and sales, and after-sales service. There are also support activities related to the infrastructure of the business, human resource management, and technology. These activities create a product that has a value to the marketplace that is determined by what customers are willing to pay for it. The difference between the value of the product and the costs associated with all the activities required to produce that value represents the margin or profit for the business. Value chain analysis deals with how to maximize a firm's profit by studying the costs associated with the various activities of the business, and reducing or eliminating costs that do not add to the value of the product (Noble, 1999).

You can view a supply chain in the same light as Porter's value chain. It takes the structure of the value chain beyond a single organization and expands the concept into a value system. The supply chain is the system through which multiple organizations deliver their products and services to customers. These interlinked organizations,

the common purpose of which is to develop the best possible means of delivering products to customers (Poirier & Reiter, 1996), work together to reduce costs and increase profits for everyone in the chain.

The supply chain models how a firm operates within its industry segment. Its focus is on the interactions of the various entities that make up the chain. Corporations are obtaining competitive advantage through the management of these linkage strategies (Gordon & Gordon, 1996) with their suppliers and customers, and sometimes even their competitors. The linkages across the supply chain have provided opportunities to use new information technologies to manage these activities. These new technologies go far beyond concepts like electronic data interchange (EDI) and production scheduling provided by most enterprise resource planning (ERP) software packages. Businesses must invest time to explore these strategies and the information systems required to support them. The stakes are high and the goals fairly well defined, but these strategies are not easily implemented due to the complexity of this environment. In fact, implementing the technology to support SCM has been defined as the process of combining "art and science" to improve the way a supply chain delivers products to customers (Koch, 2002).

SUPPLY CHAIN MODELS

The approach to SCM can be highly variable depending on the driving function of the overall supply chain. These are generally divided into three focus areas for manufacturers: sourcing-, manufacturing-, and distribution-intensive industries (Banker, 1998), and service supply chains.

Sourcing-Intensive Supply Chains

These companies have complex products and purchase many components from their suppliers. Companies in this group include automotive and durable goods manufacturers and electronics firms. Often they are in industries with rapid life cycles, such as computer manufacturing. The SCM technologies they use need the capability of handling short life cycles and the ability to build final assemblies to specific customer orders.

Manufacturing-Intensive Supply Chains

Companies in this category have heavy investments in the assets required to manufacture their products, and for this reason, high asset utilization is a primary goal of these companies. Manufacturing execution systems integrated with the scheduling process are important SCM strategies in these cases. The goal is to optimize the use of the equipment while meeting customer demand. The demand management system must be able to produce optimized plans that follow the highs and lows of customer buying patterns. Companies in this segment include textile and bulk chemical firms.

Distribution-Intensive Supply Chains

These organizations need strong functionality in demand management and distribution planning. Advanced forecasting methods made possible by technologies such as data warehousing are useful for the analysis of the massive quantities of data generated by point-of-sale systems. Manufacturers in these industries need dynamic scheduling capabilities to meet the needs of actual customer demand. This segment includes companies that provide pharmaceuticals, food and beverages, and personal care products.

Service Supply Chains

Service SCM has become more important as the economies of most industrialized nations have become more service oriented. For example, in the United States about 80% of all workers are currently employed in the service sector (Fitzsimmons & Fitzsimmons, 2001). In service industries, the process of "delivering the service" is the product. Service supply chains focus on managing the people, projects, clients, and other resources involved in delivering these services. Supply chain management in the service sector uses technology to streamline the delivery of services and optimize the use of the resources involved in the process.

AREAS OF TECHNOLOGY

The technologies used in SCM are somewhat fragmented. Although some companies have tried to be all-inclusive in handling the various technologies required, none have been especially successful. The trend has been to take a modularized implementation approach to these solution areas, often using software products from multiple vendors. Even when an all-inclusive integrated system is selected, it is most often implemented in a modular fashion. These multiple technologies are then assembled together, to form a complete SCM solution.

Technology products tend to focus in one of four primary areas: supply chain design and simulation, supply chain optimization, supply chain collaboration, and supply chain integration. Keys to successfully implementing these technologies include (Cassis, 1997): (a) Developing and refining the supply chain model until the system optimizes the behavior of the actual production environment and (b) Integrating supply chain technologies with existing legacy systems.

Design and Analysis

These technology packages include software to model and simulate the supply chain. Their purpose is to allow designers to build models of the supply chain and then simulate various activities and volume levels to determine how the modeled supply chain will perform. They are invaluable tools for predicting the performance of new supply chains and for assessing changes proposed for existing supply chains.

Simulation

Simulation of the supply chain is done using a computer model that is designed based on the characteristics of the real (or proposed) supply chain. It has the ability to generate random variations in the supply chain, just as events have a probability of occurring to a real supply chain. It

is used to analyze and plan the structure of the supply chain based on real-world variability and dynamics over time. The object is to be able to predict the performance of a given supply chain before investments are made in the physical assets needed to operate the supply chain. A well-thought-out simulation model can provide information not attainable any other way than actually building the supply chain itself.

Optimization

This group of technology products attempts to improve the performance of the supply chain by applying various optimization techniques to the activities associated with SCM. These techniques are applied in two major areas: in design and analysis activities to optimize the structure of the supply chain and in advanced planning & scheduling (APS) technologies to optimize the flow of product along the supply chain. These techniques include logic to recognize and deal with constraints that affect the execution of the supply chain and mathematical algorithms that search for optimized solutions based on the variables and constraints of the problem. More detailed information on optimization techniques is provided later in this chapter.

Active Collaboration

Technologies in this area focus on the communication and interaction of the various members of the supply chain. These tools go beyond traditional EDI software by allowing real-time communication and visibility of product levels and production schedules throughout the supply chain. Software products that provide groupware capabilities are included in this category. The need for active collaboration, sharing information, and planning data to improve the performance of the supply chain (and its individual members) has triggered a reconsideration of the supply chain model.

The traditional communication structure for members of a supply chain has been linear and synchronous:

$$\text{Supplies} \Leftrightarrow \text{Manufacturing} \Leftrightarrow \text{Distributor} \Leftrightarrow \text{Retailer} \Leftrightarrow \text{Customer}$$

The Internet has opened the door for a more collaborative communication structure for all members of the supply chain, a structure that is nonlinear and asynchronous (Figure 2). This change in the communication structure has affected not only the flow of information, but in some cases the basic structure of the supply chain itself. When this happens, some members of the supply chain may be squeezed out of the chain, such as when the Internet replaces part of the traditional distribution channel.

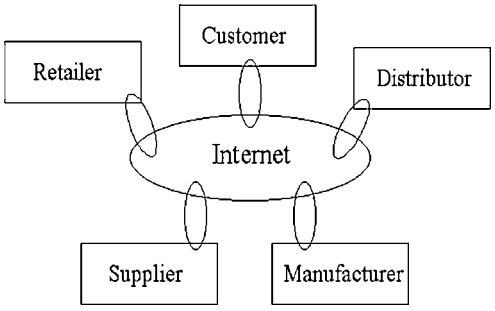


Figure 2: Collaborative supply chain structure.

Active collaboration assists the transition of supply chain members to a more customer-driven focus through the sharing of information and visibility of customer wants and needs. It also facilitates synchronized planning activities (along with APS) to help reduce overall costs through reductions in inventories between the members of the supply chain.

Integration

Supply chain integration technologies are of major concern as most companies opt for a modular approach to supply chain technologies. Various technologies are selected based on what a company wants to achieve in managing the chain. These technologies also need to be integrated with legacy applications that already exist in a business. Supply chain integration deals with linking and interfacing new technologies with legacy technologies to implement a complete SCM solution. Software products that specialize in this integration include middleware and database applications. Middleware operates by providing an infrastructure that allows all applications to communicate in a standard way. Database integration is done through the exchange of information in the database.

SCM PLANNING PROCESS

SCM planning activities fall into two categories: strategic planning and tactical planning. Strategic planning involves the top level of the planning process. At this level, decisions regarding the number of manufacturing plants and distribution centers are determined. The location of these facilities and the proximity of customers and raw materials and other components are also analyzed. To accomplish this level of planning, design and analysis and optimization tools are used. Supply chain optimization tools help determine the optimal supply chain network configuration. Supply chain simulation is used with these optimized networks to determine how various configurations will perform in different economic environments.

Tactical supply chain planning involves determining what, when, and how much each plant will produce. Optimization tools are used to generate optimal production, supply, and distribution schedules for product demand over a given time period. This planning includes details regarding the procurement and delivery of raw materials, production schedules for manufacturing facilities and assets, and the movement of finished goods through the distribution system. Some specific planning activities performed in managing the supply chain are reviewed next.

Demand Planning

Demand planning has taken on new importance in SCM as emphasis has shifted from manufacturing-centric management to customer-centric management. This paradigm shift has caused a basic change in the criteria for what

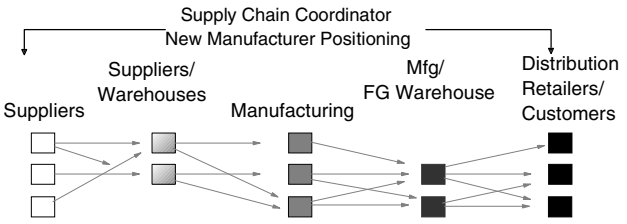


Figure 3: Positioning to control the supply chain.

constitutes demand. The focus shifts from a build-to-stock mentality, in which managers attempt to maximize the utilization of production assets, to a build-to-order mentality, in which customer needs are the top priority. In effect, the customer is scheduling the factory.

The objective becomes flexibility in the manufacturing area so that production can be scheduled based on real customer orders, instead of forecasts. To achieve this flexibility, more communication is needed downstream, with distributors and customers, and more control is needed upstream, with suppliers of the raw materials and other components needed in the manufacturing process. This requires a stronger approach to the way a company manages its supply chain. Instead of its former role of "just another link in the chain," the manufacturer needs to reposition itself above the chain (Arntzen & Shumway, 2002), to monitor and control product flow across the entire supply chain. By managing this correctly, the manufacturer can provide better visibility for itself and all of the other members of the chain. This new model is depicted in Figure 3.

Implementing demand planning requires several of the supply chain technologies mentioned earlier. Tools for active collaboration must be implemented to support the communication of customer requirements and to forecast product needs. Technologies supporting vendor-managed inventory can be used to keep current with information on customer consumption and buying habits. That information can then be used in the demand planning process to drive production scheduling. Optimization tools such as advanced planning and scheduling (APS) software can be used to establish an optimized production schedule and provide visibility and control across the supply chain to all affected participants. Each company's ERP system should then be integrated with the information provided by APS to schedule and control internal operations.

Data warehousing and data mining are other SCM tools often used to develop input for demand planning. These techniques analyze large amounts of data for trends in demand due to customer preferences, seasonality, and other factors (Carnahan, 1997).

Advanced Planning and Scheduling

The purpose of APS is to plan and develop an optimized production schedule to meet customer demand based on the constraints of the materials, production resources, and logistics of the supply chain. It goes beyond the capabilities of the scheduling provided by most ERP systems, by using supply chain optimization techniques that determine an optimized schedule based on the goals of management and the constraints of the system.

APS is used for both long- and short-term planning activities. Long-term planning involves planning over the entire supply chain and includes the interaction and coordination of all members of the chain. Short-term planning generally refers to scheduling at the plant floor level. In both cases, constraints and the optimized goals of the organization are considered in the planning process.

SCM Execution

These technologies cover a range of solution areas, including applications to track customers, the physical status of goods, the management of materials, and financial information across the entire supply chain. SCM execution technologies include systems for demand management, production execution, materials management, and transportation management.

Demand Management

Demand management attempts to optimize the performance of the supply chain by integrating supply and demand information. To do this, several supply chain technologies are used. As mentioned earlier, forecasting applications are of primary importance, especially in collaboration with other members of the chain. Demand management during execution focuses on weekly or monthly horizons, requiring close communication among supply chain members. The customer relationship management (CRM) system must also be considered, linking product requirements to other customer-related activities.

Pricing policies play an important role in managing demand as well. Seasonal or sales promotions can cause spikes in demand that are difficult to manage, especially if those further up the chain aren't aware of them. Collaborative forecasting systems link production and distribution channels with forecasters, improving forecasting accuracy by helping to determine the impact of pricing and promotions, new product introductions, and intermittent or declining demand. These sophisticated systems predict the effect "causal" or "event-driven" factors have on demand.

Demand management must provide closed-loop integration with other supply chain systems, providing accurate information for supplier, production, distribution, and customer management functions. Working together, these systems have the potential to increase inventory turns, reduce inventory obsolescence, and increase revenues for all members of the supply chain.

Manufacturing Execution Systems

Manufacturing execution systems (MESs) enable the optimization of production activities by managing the information required to launch and complete products in the manufacturing plant. Production requirements from the supply chain are communicated via the ERP system to an MES, which controls and reports on plant activities as they occur. In turn, the MES provides information about production activities to the enterprise and other members of the supply chain. Using automated data collection technologies, MES tracks work-in-process, labor reporting, and production reporting activities. By monitoring

real-time information about setups, run times, throughputs, and yields, managers are better able to identify bottlenecks, analyze performance constraints, and make adjustments to ensure that production requirements are met. This real-time response to the dynamic conditions on the plant floor results in better utilization of inventories and production assets and better delivery performance to customers. Linking the plant floor with the ERP system, MES can deliver real-time order status information to the entire supply chain.

Supplier Systems

Supplier management systems provide integration between a company and its suppliers. In effective SCM systems, suppliers must be carefully selected, because they become "partners" as members of the supply chain. Here again, collaboration is of utmost importance; information sharing and trust must be exchanged to make the process work. Through supplier databases and business-to-business purchasing technologies, companies "integrate" with important suppliers. The supplier has a view of the company's production requirements and can therefore better predict the components they must supply. This streamlines the purchase order management function, improves delivery performance, and helps reduce product costs.

Logistics

Logistics is the process of planning and controlling the flow and storage of goods from point of origin to point of consumption. It works hand in hand with the inventory management system and is often outsourced to a service provider who becomes another member of the supply chain. Logistics providers fall into the following categories:

- Transporter—The provider moves the material, but the company decides which materials, when to move them, and where to move them.
- Third-party logistics—The provider moves the material and provides storage and handling facilities between the company and its customers. The company still decides what and when.
- Fourth-party logistics—The provider takes over some control of delivery scheduling and stocking levels, as well as some of the operational functions within the company's facilities.

The more the logistics provider does, the more it needs to participate in the technologies supporting the supply chain.

OPTIMIZATION APPROACHES

Two optimization approaches, developed by operations research and the management sciences, have been incorporated into the technology toolbox of SCM. They fall into the categories of optimal solutions and heuristic solutions, and both are proving to be beneficial to the planning and execution activities of SCM. Optimization methods improve the planning capabilities of SCM by developing plans that are not only feasible (they meet demand needs and supply limits), but that are also optimal (at lowest cost, greatest profitability, or both). The need for realistic, optimized plans has shifted the material-planning task from material requirements planning (MRP), which does not consider supply constraints (Lapide, 1998), to APS systems that develop optimized solutions, based on those constraints.

Although optimization methods have been available for some time, they have become more popular in recent years as computer technology required to do the calculations involved in these techniques has become more powerful and less expensive. These methods determine solutions to problems involving limited resources by stating the problem in terms of objectives and constraints. The objectives of SCM optimization can vary but generally include combinations of profit maximization, service level attainment, cost minimization, and delivery performance. SCM optimization deals with determining what products to manufacture given the demand for the products and the constraints related to the production of those products: raw material and component supply, time, production asset availability, storage availability, and the costs of all of these components. The ideal solution will be feasible, optimized, and beneficial to all members of the supply chain.

Of course, sometimes there are no feasible solutions to a SCM problem. In this case, the optimizer will choose the least detrimental infeasible solutions. An example is when a manufacturing firm cannot produce all of the products required to fill customer orders within a given time period. When this happens, some products are placed on a back-order status and the customer orders for those products are delayed.

Of the feasible solutions, there are generally some that are optimized, and in some cases there may even be an optimal solution. Most SCM problems are so large and complex that current optimization technologies don't allow an optimal solution to be attained. They generate a number of feasible solutions and allow the user to select one. If a feasible solution cannot be attained, these systems suggest various options to the user to loosen some of the constraints so that a feasible solution may be generated.

Optimal Techniques

Optimal techniques are based on mathematical algorithms that use a series of formulae to represent the variables and constraints of the problem, and generate a result based on the goal of maximizing or minimizing some objective function.

Linear Programming

Linear programming models are used for problems in which linear equations can be used to define the objectives and constraints of the SCM planning problem. This problem-solving method results in an optimal solution. A simple example of the type of problem this approach is used for follows.

Table 2 Cost of Producing Products A, B, and C at Plants 1 and 2

	A	B	C
Plant 1	$40	$32	$81
Plant 2	$45	$29	$75

A company with two manufacturing plants has orders for three products that can be produced at either plant. The costs of producing the products at each plant are shown in Table 2.

Plant 1 can produce 200 products per day and Plant 2 180 products per day. If orders are due at the end of the week (5 working days) for 700 units of product A, 800 units of product B, and 300 units of product C, how should the plants be scheduled to produce the required products at minimum cost?

The linear programming approach to this problem would start by establishing variables and defining an objective function. Then constraints would be identified, to which the feasible solution must adhere. In this problem those would be expressed in the following mathematical terms:

Variables: $x1$, $x2$, $x3$ represent the amount of each product (A, B, and C, respectively) that Plant 1 will produce; $x4$, $x5$, $x6$ represent the amount of each product that Plant 2 will produce

The Objective: minimize (Cost) = $40x1 + 32x2 + 81x3 + 45x4 + 29x5 + 75x6$

Constraints:

$x1 + x2 + x3 <= 1000$ (the capacity of Plant 1)

$x4 + x5 + x6 <= 900$ (the capacity of Plant 2)

$x1 + x4 = 700$ (orders for product A)

$x2 + x5 = 800$ (orders for product B)

$x3 + x6 = 300$ (orders for product C)

The LP solution to this problem produces an optimum production plan with a minimized cost of $74,300, by scheduling production as shown in Table 3.

This solution was obtained by using a linear programming application based on the Simplex algorithm located on the Internet (Hochbaum & Goldschmidt, 1998). You can also find a linear programming solver in Microsoft's Excel spreadsheet program.

To describe a typical supply chain accurately to an optimizer, one would have to include, among other things, the following variables to define objectives, constraints, and relationships: resource utilization percentages, production capacity, customer demand, inventory storage capacity, minimum inventory requirements, transportation costs, throughput capacity, and product yields.

Table 3 Optimal Production Plan for Plants 1 and 2

	UNITS OF PRODUCT		
	A	B	C
Plant 1	700	200	0
Plant 2	0	600	300

Integer and Mixed Integer Programming

These programming solutions are similar to linear programming problems, except that an integer-programming problem is one in which the decision variables are constrained to integer values, and a mixed-integer problem is one that contains both integer and noninteger decision variables. In a problem similar to the one described earlier, the linear programming model might have given a production rate of 700.4 units of Product A in Plant 1. Most people would round that to 700 and move on. Suppose, however, that we were using the linear programming technique to decide how many warehouses we should build and the result was 0.4 warehouse at one location and 0.6 warehouse at another. In this case, we would want the results in integer values.

Heuristic Solutions

Heuristic solutions for addressing SCM planning are based on various models, rules of thumb, and the concept of intelligent trial and error to improve the solution results. Beginning with a known feasible solution to the problem, the heuristic approach follows rules to incrementally modify the variables in the problem and then analyze the results (feedback). If the objective improves, the process is repeated. This continues until the objective no longer gets better. This approach will produce an optimized solution, but not necessarily the optimal solution.

There are a number of heuristic methods used in supply chain planning software products based on both proprietary and published approaches. One popular method is the theory of constraints by Eli Goldratt. This method focuses on the most constrained variables, or critical bottlenecks, in addressing the planning function. The master plan (solution) is developed around these bottlenecks.

Genetic Algorithms

Genetic algorithms work well on mixed (continuous and discrete) problems. The process looks for optimized solutions from a large set of possible solutions. These are crossbred or mutated to form new sets of solutions. This continues from one generation to the next until a reasonably optimized solution is developed. Processing genetic algorithms can be computation intensive, but they generally work better than other approaches for certain types of optimization problems.

To use a genetic algorithm, one must define an objective function, the genetic representation, and genetic operators (Wall, 2002). The algorithm then creates a set of solutions and applies genetic operators such as mutation and crossover to evolve the solutions to find an optimized one.

Tabu Search

This process for finding optimized solutions is described as a meta-heuristic superimposed on another heuristic and uses methods to forbid searches that take the solution

to alternatives previously visited (considered "tabu"). It also allows the search to cross boundaries of feasibility by systematically imposing or releasing constraints to explore solutions that may reside outside local optimality. It has proven to be useful in several areas of resource planning and design for SCM. More information can be obtained on this topic from the book *Tabu Search* (Glover & Laguna, 1997).

Simulated Annealing

This technique for optimizing solutions is based on a theory from statistical mechanics. It works by simulating the process nature performs in optimizing the energy of a crystalline solid, when it is annealed to remove defects in its atomic arrangement. It is used to approximate the solution of very large optimization problems and works well with nonlinear objectives and arbitrary constraints. One criticism, however, is that it can be slow in determining an optimal solution. More information on this technique can be found in *Simulated Annealing and Boltzmann Machines: A Stochastic Approach to Combinatorial Optimization and Neural Computing* (Aarts & Korst, 1989).

Exhaustive Enumeration

This process involves looking at all possible alternatives to find the best solution and is therefore only used when there are few variable alternatives to evaluate. It takes an enormous amount of computing power to evaluate all possible combinations of a complex problem. Most SCM problems are too large for this approach. Even with today's super computers, they cannot be completely enumerated in polynomial time.

IT INFRASTRUCTURE

IT infrastructure is made up of the underlying technology products on which the SCM technologies run. It consists of the networks, databases, and application interfaces that support supply chain information flows and that allow supply chain applications to communicate with other applications. These building blocks provide the key to being able to take advantage of the many supply chain products and services available in the marketplace.

Traditionally, a point-to-point integration structure (Figure 4) has been used to link applications together. Although there may be performance advantages in the execution of applications using this approach, this becomes cumbersome and expensive to maintain in an environment where there are many components in the application portfolio.

An alternative to point-to-point integration is to develop an integration infrastructure (Figure 5) using a common interface technology in all applications.

A strategy for integration should be developed early in the supply chain design cycle. Once the infrastructure has been established, only SCM applications that can operate within that infrastructure should be considered for integration into the system. Two approaches are discussed below.

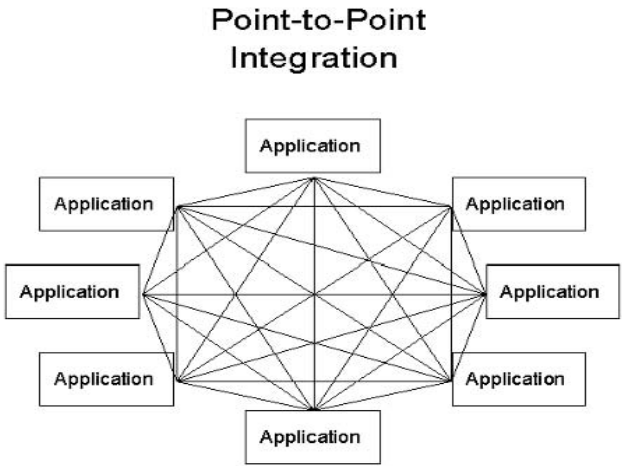


Figure 4: Point to point-integration structure.

Middleware

Enterprise middleware provides a common interface that all applications can use, no matter the operating system platform on which they run. In a message queuing environment, a set of enterprise application programming interfaces is provided for each operating system. These interfaces are invoked by the applications to provide communication and messaging capabilities, data transformation, routing, and application connections. Instead of communicating directly with other applications, each application sends information through a "hub and portal" (Arntzen & Shumway, 2002) directed by the message queuing controller. The advantage of this approach is twofold. First, all applications communicate with each other in the same way. Second, as applications are added or replaced, the other applications they communicate with do not have to be modified. An example of this type of middleware is IBM's MQ Series product set.

Database Interface

A database interface strategy uses a common database management system to facilitate the transfer of data between application systems. In this case, an enterprise

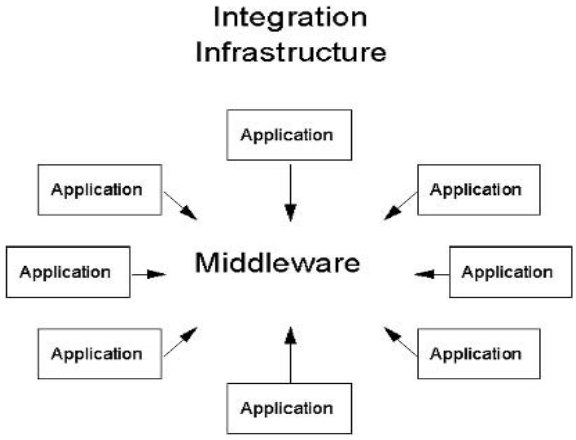


Figure 5: An integration infrastructure.

database management system is required so that applications in different operating environments can participate. Although not as robust as the message queue approach (it generally does not facilitate real-time interaction), it does provide a degree of application integration without having to interface each application to the other individually.

NETWORKS

There are three levels of networks used to support supply chain communication (Turcotte, Silveri, & Jobson, 1998): intranets, extranets, and the Internet. An intranet is used within the boundaries of a company. It contains sensitive information restricted to those within the organization. Proprietary systems, such as ERP and CRM, and the data to run them are found on intranets. An extranet is an external intranet shared by two or more companies. Data shared among companies within a supply chain is protected form the outside world on the extranet. This is where many active collaboration activities are conducted. The Internet is the public network we all know and use. Data the supply chain members want to share with the outside world is routed here.

DATABASES

Databases are an integral part of any information system. In the supply chain environment, the database management system should be standardized for all applications. It should also be an "enterprise class" database, that is, one that can operate on multiple computing platforms and is highly scalable.

APPLICATIONS

Applications involved in managing the supply chain are varied in functionality and address information-processing needs along the whole chain. Some operate exclusively within a firm, whereas others are external to the firm, most often linking members of the supply chain together. Applications such as ERP provide comprehensive sets of modules that cover a wide span of technology processing within organizations. Others, like CRM, focus on a specific area. A list of commonly used application technologies follows. Many of these have been discussed in this chapter, and more information about the others can be found in the Further Reading section.

- Advanced planning and scheduling
- Active collaboration technologies
- CRM systems
- Data mining and warehousing
- Demand management
- Distribution resource planning
- EDI
- Enterprise resource planning
- Groupware
- Logistics network design and execution
- Manufacturing execution systems
- Supply chain simulation and optimization
- Sales force automation
- Warehouse management systems
- Supply chain management solution providers

Solution providers in the SCM arena fall into three basic categories: software providers, hardware providers, and services providers. Any attempt to put together a comprehensive list of every firm involved in supply chain technologies would be fleeting at best. The most current resource for locating a company involved in a particular area of interest is the Internet. There are a number of Web sites dedicated to SCM, and virtually every SCM vendor and service provider has a presence on the Web. A fairly comprehensive list of companies and what they provide can be accessed at www.business.com, following the links Home ⇒ Management ⇒ Operations Management ⇒ Supply Chain Management (SCM).

SUPPLY CHAIN MANAGEMENT TRENDS

SCM technologies have evolved from internally focused applications to widely dispersed external systems with many participants. Communication and visibility are the new challenges as members seek to collaborate across the entire supply chain. Every member of the chain needs information from the other members. Information regarding customer orders, forecasts, and component availability must flow up and down the chain. Input for ERP, APS, demand planning, and other systems needs to be readily available to everyone. The visibility of this information, providing the ability to monitor and control the supply chain in real time is critically important. Focus must be on total system performance for the benefit of the entire supply chain.

SCM solution offerings are moving away from tightly integrated technologies that try to do everything and trending toward modular applications that can be integrated with other products (Reddy, 2002). This reduces the risk of implementation by focusing on a narrow area of functionality. It also reduces the cost of getting started with SCM. Businesses can start small and add modules as required.

The linear supply chain model is being replaced by a network (Web-like) model. The traditional model is sequential in terms of the product and information flow (O'Brien, 2000), from one member to the next, up and down the chain. The new model allows information to be available to all members at once. It's no coincidence that the supply chain model is taking on the form of a Web. The Internet has facilitated fully linked participants, allowing the collaborative management necessary in today's environment.

The Internet has spawned e-business applications for buying, selling, product design, and collaborative planning. These technologies are playing ever more important roles in optimizing supply chain performance. This has changed the way supply chains are managed and controlled, allowing different companies within the chain to synchronize their planning activities and work together toward common goals. At the same time, new channels of distribution are emerging from the use of the Internet,

which may squeeze some players out of the chain. Companies need to be alert to the technologies that will make them either winners or losers in this new supply chain environment. Supply chain management has emerged as one of the top priority strategies in manufacturing today.

GLOSSARY

Active collaboration technologies Real-time technologies (usually Internet based) that allow the various members of a supply chain to share information and work together toward a common goal.

Advanced planning and scheduling (APS) A supply chain planning tool that goes beyond the planning and scheduling of ERP by attempting to produce an optimized plan based on variables and constraints that limit the ability to deliver the right product at the right time.

Bull-whip effect A term describing the phenomenon that occurs when inventory demand variability is affected by changes further down the supply chain. The more distant an inventory buffer is from the consumer, the more variability will occur for its demand.

Business-to-business purchasing Purchasing that allows companies to integrate with their most important suppliers, usually via the Internet. It provides the capability to streamline the processes associated with purchasing.

Customer relationship management (CRM) An information system used to manage all activities related to customers, including contacts, customer orders, information requests, and technical support.

E-business Using the Internet and other information technologies to perform business process activities.

E-commerce Buying and selling goods on the Internet, sometimes used interchangeably with e-business.

Economic order quantity An order quantity algorithm to calculate how much of an item to purchase or manufacture at one time. It attempts to minimize the costs of acquiring and carrying inventory.

Enterprise Resource Planning (ERP) An integrated information system used to manage the corporate and business location processes within a manufacturing environment. Such transaction-based systems serve all functions within an enterprise and are often the data source of the decision support systems of supply chain management. Most ERP systems include software modules for manufacturing, inventory management, order processing, accounting, purchasing, and warehousing.

Extranet Two or more intranets linked together, allowing multiple companies to exchange information and share resources for a common purpose. Members of a supply chain often form extranets with each other.

Genetic algorithm A problem-solving algorithm that establishes sets of possible solutions and then uses an evolution-like process to determine the best or optimum solution.

Heuristics Problem-solving techniques that involve the use of subjective knowledge, trial and error, and rule of thumb.

Intranet A private network (resembling the Internet) for the exclusive use of those who are authorized to use it, normally those within a company or other organization.

Linear programming A mathematical technique used to obtain an optimum solution in resource allocation problems, such as production planning.

Vendor managed inventory The arrangement by which vendors manage inventory they own inside the stores of their retailers. The retail stores become an extension of the vendor's warehouses, allowing them to track inventory at the point of consumption.

CROSS REFERENCES

See *Business-to-Business Electronic Commerce; Customer Relationship Management on the Web; Electronic Commerce and Electronic Business; Enterprise Resource Planning (ERP); Intranets; Inventory Management; Supply Chain Management*.

REFERENCES

Aarts, E., & Korst, J. (1989). *Simulated annealing and Boltzmann machines: A stochastic approach to combinatorial optimization and neural computing.* Hoboken, NJ: Wiley.

Anderson, B. F. (1997, April). The seven principles of supply chain management. *Supply Chain Management Review*. Retrieved April 23, 2002, from http://www.manufacturing.net

Arntzen, B. C., & Shumway, H. M. (2002, January). Driven by demand: A case study. *Supply Chain Management Review.* Retrieved April 23, 2002, from http://www.manufacturing.net/scm/index.asp?layout = articleWebzine&articleid = CA197691

Banker, S. (1998, August). Different supply chains require different APS solutions. *Advanced Planning & Scheduling Magazine,* pp. 12–19.

Bendiner, J. (1998, January). Understanding supply chain optimization: From: what if to what's best. *APICS The Performance Magazine,* pp. 34–38.

Carnahan, M. (1997, November). Data-powered decisions making. *APICS The Performance Advantage,* pp. 34–36.

Cassis, S. (1997, November). Understanding advanced planning systems. *APICS The Performance Advantage,* pp. 30–32.

Fitzsimmons, J. A., & Fitzsimmons, M. J. (2001). *Service management: Operations, strategy, and information technology* (3rd ed., p. 5). New York: McGraw-Hill.

Glover, F., & Laguna, M. (1997). *Tabu Search.* New York: Kluwer Academic.

Gordon, S. R., & Gordon, J. R. (1996). *Information systems: A management approach* (p. 82). Orlando, FL: Dryden Press, Harcourt Brace & Company.

Hochbaum, D. S., & Goldschmidt, O. (1998). LP Solver. Retrieved August 2, 2002, from http://riot.ieor.berkeley.edu/riot/Applications/SimplexDemo/Simplex.html

Koch, C. (2002, January). The ABCs of supply chain management. *CIO.com.* Retrieved April 23, 2002, from http://www.cio.com/research/scm/articles

Lapide, L. (1998, May). Supply chain planning optimization: Just the facts. *Advanced Manufacturing Research,* pp. 3–30.

Lee, H. L., Padmanabhan, V., & Seungjin, W. (1997). The bullwhip effect in supply chains. *Sloan Management Review, 38*(3), 93–102.

Noble, H. (1999, October). Key enablers for supply chain management. *APICS the Performance Advantage,* pp. 60–62.

O'Brien, K. P (2000, April). Value-chain report. *Industry Week's The Value Chain. IndustryWeek.com.* Retrieved April 23, 2002, from http://www.iwvaluechain.com/columns/columns.asp?columnid = 598

Poirier, C. C., & Reiter, S. E. (1996). *Supply chain optimization.* San Francisco: Berrett-Koehler.

Porter, M. E. (1985). *Competitive advantage: Creating and sustaining superior performance.* New York: Free Press.

Reddy, R. (2002, January). The evolution of supply chain technologies. *Intelligent Enterprise.* Retrieved April 23, 2002, from http://www.intelligententerprise.com/020114/502infosc1_1.shtml

Turcotte, J., Silveri, B., & Jobson, T. (1998, August). Are you ready for the e-supply chain? *APICS The Performance Advantage Magazine,* pp. 56–59.

Wall, M. (2002). Introduction to genetic algorithms. Retrieved April 23, 2002, from http://lancet.mit.edu/~mbwall/presentations/IntroToGAs/

FURTHER READING

For those seeking a more comprehensive understanding of this topic, the following textbooks on supply chain management are recommended. There are also many other resources available on the Internet under the search words "supply chain management."

Chopra, S., & Meindl, P. (2001). *Supply chain management: Strategy, planning and operation* (1st ed.). Upper Saddle River, NJ: Prentice-Hall.

Fitzsimmons, J. A., & Fitzsimmons, M. J. (2001). *Service management: Operations, strategy, and information technology* (3rd ed.). New York: McGraw-Hill.

Govil, M., & Proth, J.-M. (2002). *Supply chain design and management: Strategic and tactical perspectives.* San Diego, CA: Academic Press.

Monczka, R., Trent, R., & Handfield, R. (2002). *Purchasing and supply chain management* (2nd ed.). Mason, OH: South-Western Thomson.

Shapiro, J. (2001). *Modeling the Supply Chain.,"* Pacific Grove, CA: Brooks/Cole-Thomson Learning.

Supply Networks: Developing and Maintaining Relationships and Strategies

Robert H. Lowson, *University of East Anglia, United Kingdom*

INTRODUCTION

The world is changing. The consumer is spoiled. Diversity is rampant. We have moved away from the supply side of business to a"pull" world. Consumer demand is approaching the chaotic in its insatiable appetite for diverse, individualized services and goods that are provided by flexible and responsive organizations. To understand this shift, management theorists have developed a whole galaxy of operations strategies and operational activities, including supply network management. For many firms competing in increasingly complex and dynamic sectors, the correct choice, implementation and evolution of a supply network operations strategy, can provide considerable competitive advantage. Pivotal to such strategies are the core relationships involved.

DEMAND COMPLEXITY AND THE NEED FOR A FLEXIBLE RESPONSE

Demand tells us exactly what products, services, and value customers seek. In other words, as Lauterborn (1990) suggests, identifying and delivering the four C's: customer needs and wants; cost to the customer; convenience; and communication.

This chapter examines the nature of supply network relationships and their contribution in satisfying the demands of the market and customer. Before embarking on this task, however, we need to spend a little time thinking about the concept of flexibility. It is a subject that currently exercises the minds of many managers in today's organizations, as flexibility, in its various guises, offers the potential to answer the increasingly heterogeneous demands of the customer. Flexibility can be examined from two particular perspectives: the stimuli for greater product and service variety, and the various classifications of flexibility organizations exhibit by way of a response.

The Stimuli for Greater Flexibility

Flexibility in operational systems is a response to the need for variety and its attendant uncertainty. The former can be viewed as being demand driven whereas the latter is a supply dilemma. We examine both of these elements in a little more detail.

Consumer Demand for Variety

Consumer tastes in the past four decades have altered radically. Consumer demand in many sectors for both goods and services is displaying piecemeal, disjointed, and unsystematic tendencies and, thus, is becoming increasingly difficult to satisfy. Consumer purchases are more than ever a reflection of a lifestyle or fashion statement rather than the satisfaction of a basic need, and this is only the beginning. To this has to be added the complexity of instantaneous, electronic, worldwide communication, and an information explosion that has served to educate the average shopper beyond any level seen to date. To meet these changing demand patterns, organizations are having to react more speedily while avoiding the penalties associated with increasingly volatile demand. In many sectors, this drive for variety has reached almost chaotic proportions.

Empirical research confirms some of these trends. In a recent survey of U.K. retailers in fast-moving consumer goods (FMCG) industries, 85% of respondents agreed that "for-sale" seasons were becoming much shorter. In addition, 93% of the survey witnessed a trend toward greater product volatility, fashion influence, difference, and customization over the past 3 years. Similar questions were posed in a separate survey of supplier manufacturers. Here, 81% agreed that seasons were much shorter, 87% saw more volatility in style introductions, and 93% believed unique and customized goods to be of increasing importance (Lowson, King, & Hunter, 1999).

These changes in the business environment bring about an unprecedented reconfiguration of consumer demand. In turn, these forces are being passed along supply and demand systems. Unfortunately, the farther away from the point of sale, and from any indication of demand preference, the harder it becomes to react in an effective and efficient manner. And certainly, recourse to long-term forecasting becomes virtually useless as complexity and dynamism continues to grow. This brings us to the second problematical area when considering the driving forces for variety: the inherent uncertainty this brings.

Demand Uncertainty

If we accept the picture of increasing variety in both goods and services and the search for ways to customize products at an individual level, we can see that many industries are characterized by complexity and dynamism. Unfortunately, modern organizations are ill-equipped to deal with such uncertainty. A short history lesson:

At the turn of the 20th century, manufacturing, for example, was characterized by an emphasis on mass markets, high volume, and the use of interchangeable parts. When the principles of scientific management, as promulgated by Frederick Taylor and his disciples, were also adopted, it produced a new era of industrial power that was eagerly exploited by the likes of Henry Ford, Isaac Singer, and Andrew Carnegie.

The dogma was clear. For utmost efficiency in any factory, the following was advised: divide work into the smallest possible components; assign the tasks to specialists; appoint managers to supervise and make decisions, leaving workers free to concentrate on manual tasks; reduce variation to a minimum; standardize all inputs and outputs to reduce defects; exercise control through a rigid hierarchy that channels communication in the form of exception reports upward and directives downward; measure performance by cost, scale, experience, and length of production run; and employ forecasting systems to anticipate any possible changes.

Then in 1974, Wickham Skinner proposed the idea that manufacturers focus their plants (or even in departments within plants) on a limited range of technologies, volumes, markets, and products, and that strategies, tactics, and services be arranged to support that focus. The maxim was that a factory that succeeds in focusing its activities will outperform one that does not. Costs would be lower than in unfocused operations due to experience curve and scale benefits; consequently, focus would provide competitive advantage. Today, these philosophies continue to be espoused under the heading of lean production and lean thinking (although their supporters would have us believe that lean also means agile).

There are, however, always trade-offs with such approaches; for example, low cost and flexibility are incompatible bedfellows. If the market demands greater variety and diversification, the focused factory comes under considerable strain, often alleviated only at the expense of high inventory levels.

As we reached the 1980s, it soon became apparent that organizations operating in this manner were unable to cope with one particular demand: variety. Fundamental and radical new methods of organization and management were needed once the demand for diversity reached a critical level. We are still searching for many of these new approaches (many of which we now call operations strategies).

The next factor in the flexibility debate is the changes that these stimuli have engendered in organizational activity.

The Commercial Response to Changing Consumer Demand

Today's competitive environment requires enterprises to seek greater product and process variation (flexibility) through agility and responsiveness, increasingly rejecting the Fordist principles of mass manufacturing and moving toward mass customization.

However, this "Holy Grail" of mass customization can only be reached once an entire supply network (supply networks' or 'supply systems' are our preferred terminology to describe what is often known as the 'supply chain'), is integrated and data openly shared for visibility at all stages (Shore, 2001). A supply system, or perhaps demand system, that is fully visible provides the basis for flexibility and responsiveness to real-time demand.

How close are we to achieving such a situation? Figure 1 shows the evolution of such a supply system. Stage I depicts the situation for most goods and services sectors at the turn of the past century. The participants considered themselves separate and autonomous industries with no influence, reliance, or interconnection with each other. The second stage (II) shows the effect of interdependence, and the supply chain is born with all its accompanying opportunities and pitfalls (Lee & Billington, 1992). However, industries remained manufacturing and product driven. In Stage III we see the current situation (theoretically), with electronic data interchange (EDI) beginning to form an interface between industries in a pipeline responding to consumer pull.

In the light of the changing business environment described above, a fourth, more radical stage is needed for successful competition and the necessary flexibility. Figure 2 shows Stage IV, what we term the "cluster of value." This is a future vision. The consumer group is the nucleus of activity, dictating and driving all demand preferences for variety. The entities circling (sales processes, operations processes, vendor processes) are all closely linked in a network relationship and respond by providing value when, and in the exact format, required. Consumer groups are constantly changing over time, and, an important aspect, organizations will be involved in many different simultaneous clusters of value (an issue to which we will return). This type of structure and interaction is necessary in response to a growing demand for variety and moves toward mass customization.

A Classification of Flexibility

Operational flexibility can be witnessed at three levels in an organization: (a) At an inter- and intra-organizational level—the strategic choices made by the firm and its supply systems concerning the ability to offer a particular level of flexibility in products or services; (b) at an operations level—whether in a distinct operations

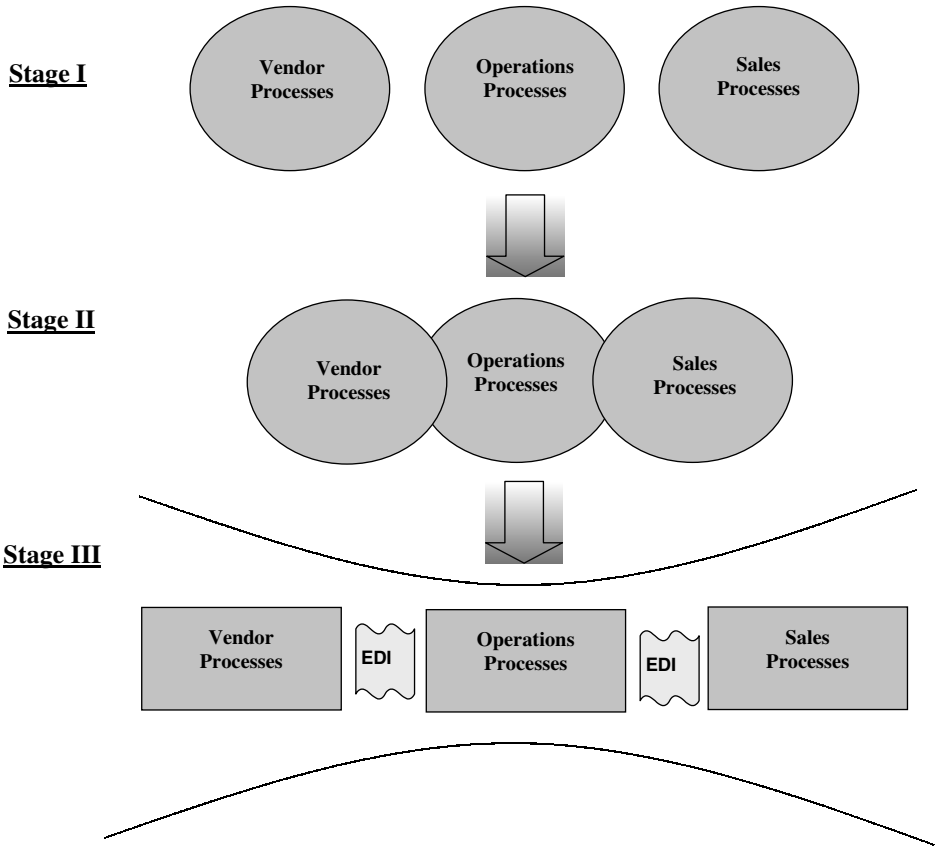


Figure 1: Supply system evolution.

function or throughout the organization, a concern for whether the operations activities are capable of sufficient product/service flexibility; and (c) at an individual, resource, process, and structure level—a concern for whether human resources, other resources (machinery, etc.). processes (stages or activities necessary to complete a task), and structures (how the system is organized and governed) are sufficiently flexible to match the variety of tasks required to support levels (a) and (b)? The strategic nature of flexibility applies at total supply network, organizational, and operational levels [levels (a) and (b) above]. Thus, we can conceptually think of flexibility as having demands that elicit a particular organizational responses These can have both external and internal implications for providing customer value needs.

An External Response to Customer Value Needs

How the firm initiates changes that have an external impact on demands for flexibility. The organization will be concerned with its ability to offer the following:

Product or service flexibility: The ability to introduce and modify products and services according to demand variety.

Mix flexibility: An ability to change the range of products or services produced over a given time period.

Volume flexibility: Being able to change the level of output over time.

Logistics flexibility: An ability to provide the flexibility that determines when, where, and how a product or service is provided.

Monetary flexibility: A flexibility over when, where, and how payment is made for goods and services (including interests rates, installments, etc).

Contact flexibility: The degree of direct contact with customers and the influence this has on value and customer satisfaction.

Flexibility also requires a firm's internal boundaries to be redrawn. Atkinson (1984), using the flexible-firm model, debated the implications for manpower, work, and employment. He identified three types of internal flexibility, to which a further two have been added.

An Internal Response to Customer Value Needs

We can think of flexibility as also having an internal impact.

Functional flexibility: Redeployment of workers as production tasks require.

Numerical flexibility: The capacity to make changes in employment levels to match demand.

Financial flexibility: Pay and other employment costs reflecting the objectives of numerical and functional flexibility.

To these we add the following:

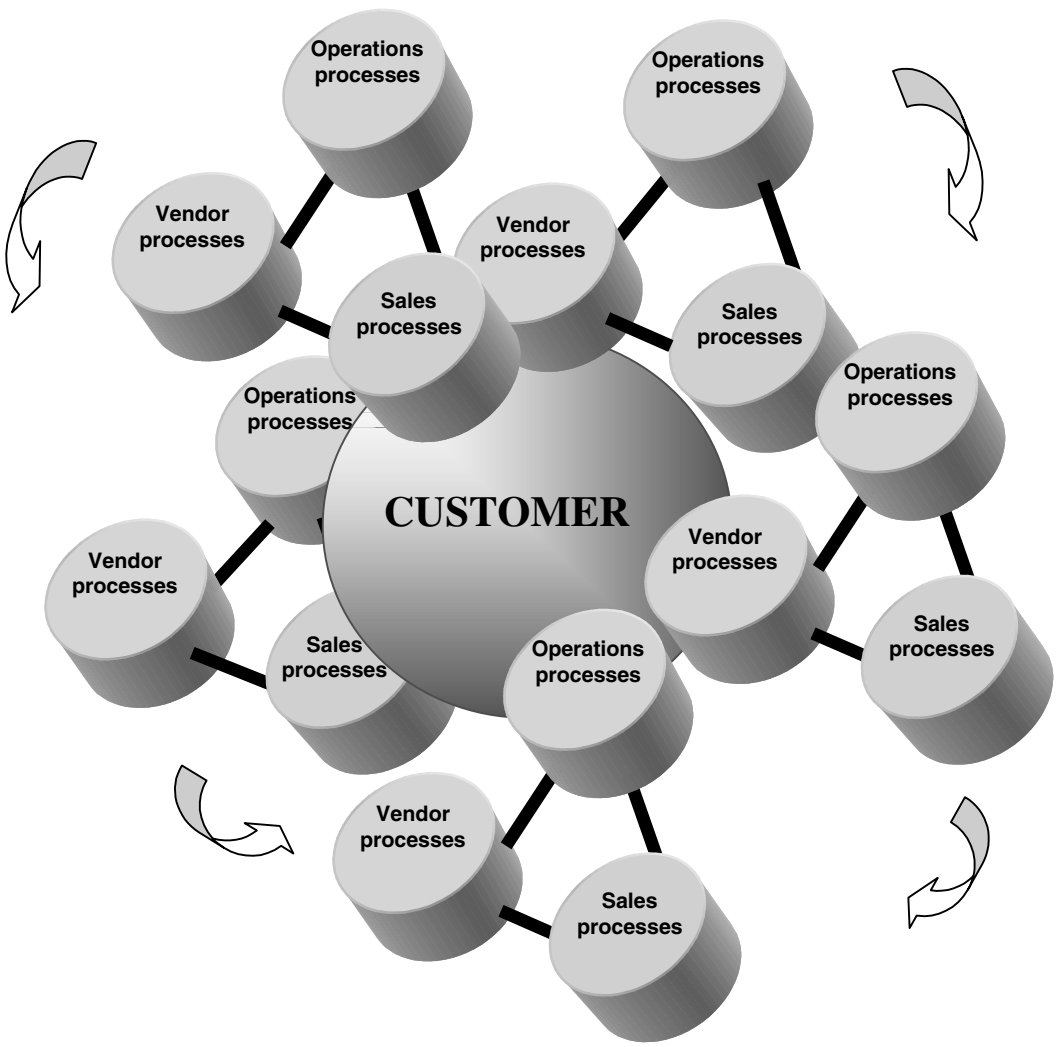


Figure 2: A cluster of value.

Temporal flexibility: Flexibility in timing work arrangements, processes, and activities to match demand flows.

Technological flexibility: The flexibility of process technology to be used for multiple purposes.

Given the importance of flexibility, it seems advisable for an organization to devise methods to ascertain the levels it requires. However, flexibility is a concept that is difficult to quantify and operationalize—to a degree, each business must assess its individual impact in terms of general benefits and costs. Closely linked to demand complexity and the flexibility debate is the notion of mass customization, and it is to this that we now turn.

MASS CUSTOMIZATION

Over the past decade, management literature has been replete with calls for mass customization [for a balanced debate, see Zipkin (2001)]. This "movement" recognizes an increasing demand for individual, customized, and personalized goods and services with the need for volume. We can see how this trend has emerged by considering the main themes witnessed in the evolution of the operations environment.

Strategic Classifications

Historically, the following strategic factions have been evident:

Mass production: This entails high volume, standardized goods and services, low variable cost, and economies of scale with little or no variety.

Lean production: As per the Toyota production system, this is the removal of all waste from the operations environment.

Mass customization: This is similar to mass production, but with variety. Instead of selecting one variety of a product, each customer provides unique information

so that the product can be tailored to her or his requirements. It requires flexible production processes and delivery capability. It is sometimes referred to as "superficial customization."

Agility and flexible operations: The agile producer aims to provide highly customized products at a cost comparable to mass production within short lead times. The tailoring of products to demand includes a higher element of service and, thus, greater added value. A flexible workforce, structure, and production technologies (especially through the use of computer-integrated manufacturing) are all contained within a learning culture. Externally, the concepts of vertical integration and long-term partnerships are replaced with short-term, flexible contracts and horizontal outsourcing, allowing rapid response through an expansive system of communication networks.

Flexible specialization: This is rejection of "Fordism" and mass production, satisfying a new demand pattern moving toward individualization. It is a return to a craft form of production, based upon the use of information technology and customized, short-run manufacture, in a network of small firms operating in niche, segmented markets.

Customized supply network operations strategies: We are now beginning to witness a new trend. Recent empirical research (Lowson, 2001a) suggests that many organizations (retailers and manufacturers) are reacting to the cluster of value scenario by adopting more than one type of supply network operations strategy. These strategies include the development and maintenance of strong supply network relationships. This has three major implications. First, just as a firm may take part in more than one cluster of value, it will often also use more than one supply network operations strategy. Second, these strategies are customized to meet the individual needs of each situation, for example a particular customer or a main product group—an operations strategy for each individual demand situation. This simple yet powerful approach provides organizations with the ability to match and respond to the demand complexity of the value stream using unique operations strategies. Third, the operations strategies deployed are intended as integrative relationship devices in a supply network, and as such, they are broader in scope than a functional strategy, such as that for manufacturing or production. We return to these issues later.

DEVELOPING AND MAINTAINING SUPPLY NETWORK RELATIONSHIPS

Against this background of customized supply network strategies, we can examine necessary relationships and their development and maintenance. First, however, a few misconceptions need to be addressed.

Genealogy and Definitions

The broad subject of supply networks, supply chain management, demand chains, strategic purchasing and supplier development, quick response, and so forth, has been one of the most contentious areas of management theory. There is a certain amount of confusion as to what does and does not constitute a supply network or chain and what their attendant operations strategies are. Indeed, some commentators are of the opinion that a supply strategy subsumes all these terms (Harland, Lamming, & Cousins, 1999). Clarification is in order, and to do this we propose a genealogy as an overall guide (see Table 1).

Table 1 A Genealogy of Supply Network Operations Strategies and Approaches

Classification	Description
Kingdom (the highest category of taxonomic classification)	Organizational business strategies.
Division or phylum (a generic group)	A generic model of a supply network operations strategy that is not any particular identifiable type, but containing interrelated building blocks.
Class (a grouping of organisms)	An identifiable type of supply network operations strategy (quick response, efficient consumer response, etc.) demonstrated in a qualitative pattern of organization. This will then be physically embodied in an individual and quantifiable deployment (the structure) unique to each situation.
Subclass	A narrower operations strategy used in a linear supply chain, value chain, or part of the chain (logistics strategy for example).
Order (taxonomic rank constituting a distinct group)	Strategic decisions about the various building blocks of a supply network operations strategy (the order).
Genus (taxonomic grouping containing several species)	Groups of building blocks (or species) form a particular operational or tactical approach, such as supply chain management and logistics.
Species (individuals with common characteristics—in practice the species will be made up from subspecies or elements)	Individual building blocks are the species: core competencies, capabilities and processes; resources; technologies; and certain key tactical activities that are vital to support a particular strategy or unique positioning. These building blocks are grouped into a class of supply network operations strategy (a specific instance) or described in the generic form (the phylum). They will also be used at a more tactical level as a particular operational approach (the genus).

We can begin at the bottom of the table with a distinct *species*. These are the generic "building blocks" of an operations strategy for a supply network or a tactical supply network approach. The building blocks will, in practice, contain certain discernable elements when deployed (*subspecies*). Next, groups of building blocks (species) form a particular operational or tactical focus (*genus*). Strategic decisions are then made about these building blocks (*order*) as components of an operations strategy. A particular identifiable type of operations strategy (*class)* constitutes the pattern of organization or form that is a qualitative set of relations. Thus, the *class* is an identifiable type of operations strategy for a supply network. This pattern or organization is then physically embodied in the structure, a unique, individual, and quantifiable operations strategy evident or used in a particular supply network situation (the physical embodiment of the class). We can also see particular *subclasses,* where the strategy is much narrower, as used, for example, in a linear chain rather than a network, or in just part of a chain (logistics). The class then forms part of a higher *division* or *phylum,* which is a generic model of an operations strategy that contains interrelated building blocks. Finally, the generic operations strategy belongs to a larger *kingdom* of business strategies.

We will return to some of these points in a moment. However, at this stage we also need to provide some clear definitions of a supply network and a supply network operations strategy. Christopher (1992) offers the following definition of a supply network: "an interconnection of organizations which relate to each other through upstream and downstream linkages between the different processes and activities that produce value in the form of products and services to the ultimate consumer."

A definition of any operations strategy for the supply network can be adopted from Lowson (2002): "major decisions about, and strategic management of, certain medium- to long-term operational capabilities. These include: core competencies, capabilities and processes; technologies; resources; relationships; and, key tactical activities necessary in any supply network; to create and deliver product and service combinations and the value demanded by a customer. The strategic role involves the blending of these various building blocks into one or more unique strategic architectures."

For completeness, we can think of supply chain management as being composed of operational or tactical activities (genus—see Table 1) and defined as "the management of upstream and downstream relationships with suppliers and customers to deliver superior customer value at less cost to the supply chain as a whole" (Christopher, 1998).

Supply Network Relationships

There are a number of possible relationships in a supply network. We can picture these as being part of a continuum that to one extreme concerns acquisitions through vertical integration and to the other, covers transactions in a pure market (see Figure 3). Thus, inter-firm relationships can vary between transactional with many vendors (very few activities performed internally) and "in-house" (few vendors).

Above the line in Figure 3 are the various relationships possible between what we might term full ownership (acquisition of a supplier or customer, placing it in-house) or purchase (transactions in the marketplace). In between are several network relationships that, moving from right

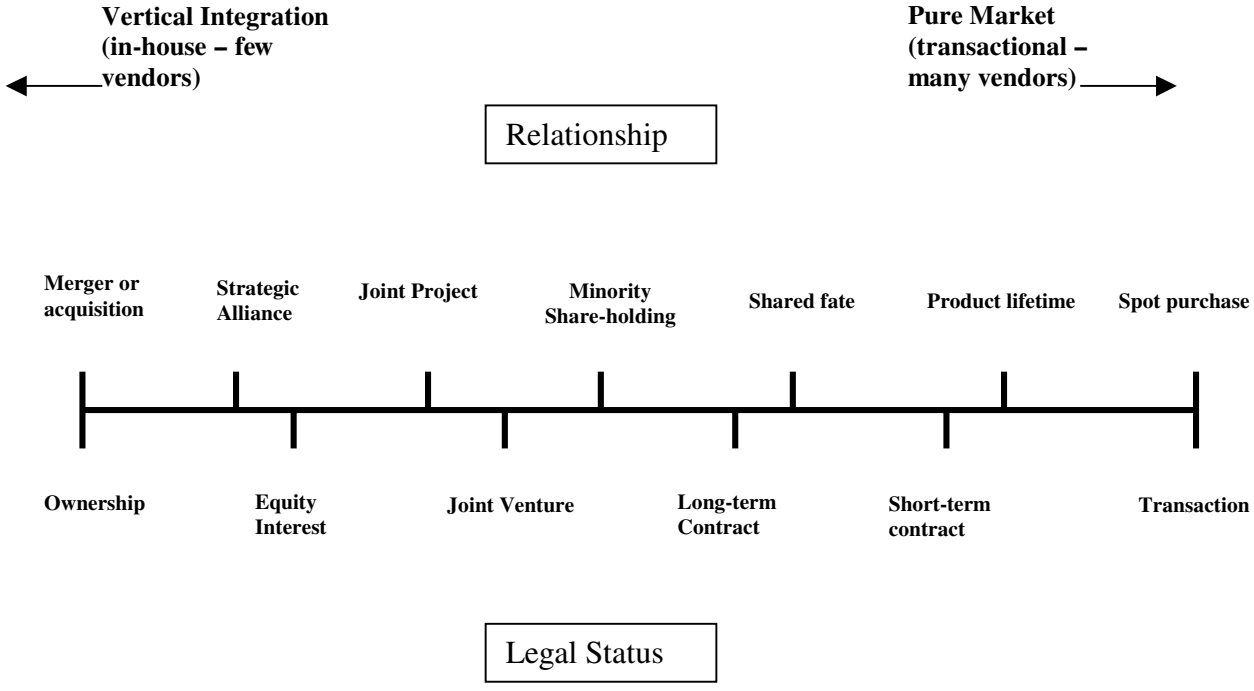


Figure 3: Relationships in a supply network.

to left (clearly, there will be many situations of overlap), have increasing degrees of commitment. Below the line the legal status can be viewed. Again, we have the situation where to one extreme we have ownership and to the other we have a market transaction. Differing arrangements are then depicted (moving from right to left) as the legal obligations become more onerous and long-term. We need to consider some of the forces driving decisions regarding whether to retain activities in-house (make) or rely on market transactions (buy).

Reasons for Adopting a Particular Network Relationship

First, management and organizational theory is subject to a number of trends—the operations management field is no exception. As we have seen, operational focus to retain scale and experience benefits was an extremely effective strategy before the need for greater variety. At this time, it made sense to retain activities in-house and to exercise control over core competencies. As demand for variety increased, firms began to realize that it was impossible to maintain expertise in everything. Better efficiency and operational effectiveness could be achieved by concentrating on a few core capabilities and outsourcing the rest to experts in that particular field. This has proved a popular strategy and one that was aggressively pursued for a number of decades. However, at the time of writing, we are beginning to see something of a reversal. It would seem that such enthusiastic outsourcing resulted in a loss of control of some vital functions, leading to higher costs. Further, it ignored the important maxim that the linkages between activities are a key source of competitive edge.

The second reason for adopting a particular network relationship concerns a reduction in transaction costs. The recent trend toward core, "preferred" vendors, in theory, reduces complexity, increases control, saves time, lowers transaction costs, and generally allows more time to devote to a joint partnership.

Third, and closely related to the above, organizations are now recognizing the value of close linkages in a supply network. These can be managed by using closer, longer term strategic alliances with fewer suppliers that can leverage higher performance in the whole system.

Advantages and Disadvantages of Relationship Types

Vertical integration describes the extent to which a firm owns the network of which it is a part and is, perhaps, the closest form of partnership relationship. Integration can be vertical or horizontal and is a strategic move to secure resources, products, or access to a market. Ownership is not always the sole reason for integration; such decisions can be taken to increase capacity and to ensure the correct network balance. Integration has a number of advantages: (a) It secures a dependable supply of goods and services, controlling sources of key materials, components, or services in-house rather than trusting the vagaries of the market. (b) It helps reduce costs. A firm may have the skills, technology, or specialist processes in-house and can utilize these at a much cheaper rate than buying such services. (c) It can improve the quality of a service or product. Specialist knowledge of a product or process is kept in-house to protect a unique innovation. (d) It helps build and advance knowledge of the market. Ownership of certain activities will help firms continue to understand demand and keep in touch with the value sought by customers.

Many firms will choose less formal supply network relationships and will tend to prefer to rely on market-based transactions due to the following disadvantages: (a) It creates a monopoly market. Customers may be faced with little choice as prices increase and service and/or quality decline. (b) It creates an inability to exploit economies of scale. Use of specialist suppliers dedicated to a particular product or process will offer economies of scale because they can produces higher volumes and, thus, lower prices. c) It creates a loss of flexibility. Heavily integrated firms will be committed to particular investments in resources and technology—they will have difficulty changing direction and offering different products and services if market forces alter. (d) It creates a lack of innovation. New product and service flexibility can be reduced, as the firm may have heavily invested in certain activities, processes, and technologies, as part of an integration move, and may find it hard, both economically and emotionally, to justify a switch in direction. In addition, the firm will also lack the creative input of suppliers and customers. (e) It creates a distraction from core capabilities. The more integrated a firm becomes, the more activities it will have to undertake—needing to be "all things to all people"—thereby neglecting its core competencies, which are a key determinant of competitive advantage.

Table 2, summarizes how different types of external supply network relationship involved differing implications.

Components and Classification of Supply Network Relationships

In general terms, a supply network relationship can be conceptualized as containing three main components (the degree of emphasis placed on each will differ, depending on the situation and the purpose of the relationship).

Strategic business components: The strategic expectations and longer term objectives sought by the relationship.

Operations strategy components: Medium- to long-term aspirations and decisions regarding the capabilities of the network at an operational level.

Operational components: The tactical, operational activities that firms in a network are required to undertake to provide products and services.

The Benefits of a Supply Network Perspective

Increasingly, especially in fast-moving consumer sectors, operations and their accompanying strategies must be viewed from a network relationship perspective. This has a number of strategic implications. The first is knowledge. Thinking at a network level encourages a broader perspective regarding the supply system. This builds strategic thinking beyond the firm itself and its immediate suppliers and customers. The interconnections,

Table 2 Implications of a Supply Network Relationship

Network Relationship	Implications and Issues
Adding value to products, services, and processes	External relationships will add value to products, processes, and services. Suppliers, customers, and other parties in a supply network offer an important input into these issues. Time to market, modifications, service delivery time, complementary product, and services offerings, etc., will all add perceived value.
Improving access to the marketplace	Network relationships offer increased market channel access and closeness to the customer. Information will, in theory, flow along seamless and transparent interfaces, reducing the distance between all organizations and the end consumer.
Improving operations	Close supply relationships improve operations by reducing costs and cycle times. Resources can be shared and utilized more effectively.
Technological development	Network relationships help to build a skill base and the better application of technology.
Competitive growth	The pooling of expertise and resources opens new market opportunities and reduces the barriers to entry.
Skills	Greater organizational learning can be achieved in a supply network relationship (covering both internal and external influences).
Financial strength	Income can be increased and financial costs shared using network relationships. Further, investment risk can be reduced by shared exposure.

amplifications in demand, end consumer, information flows, and external interdependence are brought into sharp focus. Second, the establishment of network relationships is a strategic viewpoint that inevitably moves beyond mere tactical thinking. Long-term decision making encourages the analysis of likely future-demand trends and the effects on changes in the supply network structure. This will also have implications for decisions regarding vital resources and the technologies necessary for the future.

The Partnership Process

Empirical research by Christopher and Jüttner (2000) has established that there are five key elements involved in a supply network relationship. Managers in this study were questioned as to the most important aspects of a relationship and the findings are outlined below.

Partner Selection and Classification

The differences in degree of partnering relationships are an issue in terms of the benefits to be gained, the resources and investments that are necessary and the risk involved. Relationship management is a situational approach and involves the development and maintenance of a portfolio of relationships with different natures. Thus, economic and strategic concerns are often balanced with integration needs at the operational level of the working relationship. Making the right selection to embody the necessary relationship portfolio is a crucial task.

Training of Boundary Spanners

Individuals, as boundary spanners, play a major role in establishing and maintaining inter-organizational relationships. However, the collaborative skill and organizational structure necessary to support these roles must be properly developed. This process needs to be actively managed and will increasingly apply to all employees in the enterprise.

Coordinating Interpersonal Relationships

How can the various interpersonal relationships at different organizational levels be coordinated to represent a consistent corporate approach? Relationships based upon multifaceted interface structures are extremely complex. These relationships need to be managed and institutionalized at the business or corporate level and not left to the discretionary behavior of individuals. Consequently, the commitment of senior management will have a significant affect on the success of long-term relationships. Further, having a strategic vision of such relationships is a primary means to cascade the policy down through the firm in a supportive culture.

External Support

Unfortunately, in many commercial sectors, business relationships have been primarily adversarial. The development of collaborative supply network approaches and the solving of inter-firm problems require new planning and implementation techniques. Appointing a relationship promoter is an important step in this process. Successful relationships are often developed under the auspices of a committee system involving the most important stakeholders in the relationship (internal and external to the firm).

Relationship Monitoring

A successful relationship involves the management of continual value-creating processes, underpinned by a monitoring procedure. This will complement traditional

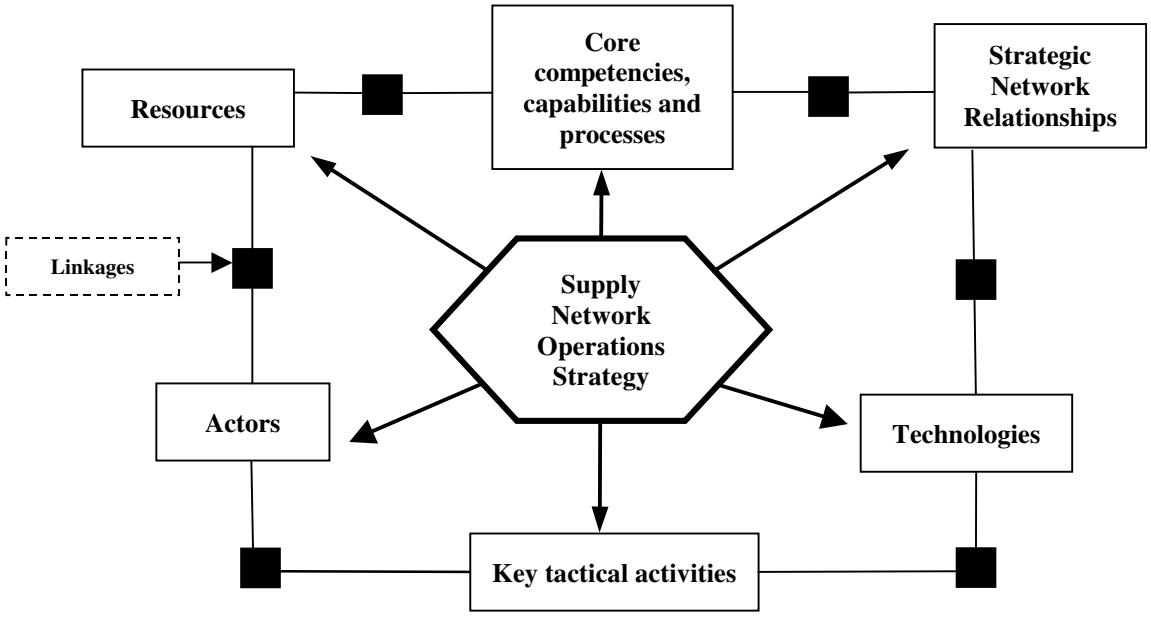


Figure 4: Components of a supply network operations strategy.

value-assessment methods. Monitoring and feedback of relationship performance and process necessitates team participation and involves all individuals working on boundary spanning activities.

Having reviewed the nature of a supply network relationship, we can now establish a connection between the earlier debate regarding demand complexity and the need for flexible response, with the types of supply network relationships likely to be seen in the future as part of an operations strategy.

CUSTOMIZING SUPPLY NETWORK RELATIONSHIPS AS PART OF AN OPERATIONS STRATEGY

In the first section, the cluster of value was introduced, with the suggestion that firms may simultaneously take part in more than one type of supply network. It seemed a logical supposition supported by research evidence, that firms will also utilize a number of operations strategies to match the complexity of their environment and achieve competitive advantage (Lowson, 2002). It was also postulated earlier in the first section, that each of these operations strategies would rely heavily on a number of building blocks; including supply network relationships. We now need to consider the process by which these strategies are developed and deployed.

The Building Blocks of a Supply Network Operations Strategy

In any supply network, an organization will view particular aspects of its undertaking as being of strategic importance. These aspects include core competencies, capabilities, and processes; strategic network relationships; resources; technologies; key tactical activities to support a particular strategy or positioning; the actors in the network; and its linkages. Figure 4 portrays these in a schematic.

The building blocks of the operations strategy will reflect the strategic significance of the supply network strategy adopted by the firm. Additionally, the firm will apply a strategy that will represent a particular "blend" of all these building blocks in order to be sure they align closely with the organization's strategic priorities. In Figure 5 we see an example of a composition matrix forming a supply network operations strategy with certain generic components. Each element in the matrix has a particular emphasis depending on the firm and industry (hence the shading), with higher order strategic themes being dark.

We now add a little more detail to these building blocks.

Core Competencies and Capabilities

A firms core competencies and capabilities, can be (a) process based (directly derived from transformation activities); for a manufacturer such as Motorola these could include the excellence of their production facilities; (b) system or coordination based (across the entire operational system), as seen, for example, in Wal-Mart, with fast replenishment of goods, vendor-managed inventory, a broad variety of products, and the ability to customize individual stores and product ranges and to modify and develop new products quickly; (c) organization based (across the entire organization), as, for example, Nordstrom and Disneyland; and (d) network based (covering the whole supply network), as in the case of FedEx and McDonald's.

Strategic Network Relationships

Strategic network relationships are the key relationships, interconnections, and associations formed over time and responsible for supplying inputs, whether services, products, components, or information. Clearly, decisions will need to be made as to whether a particular input is likely

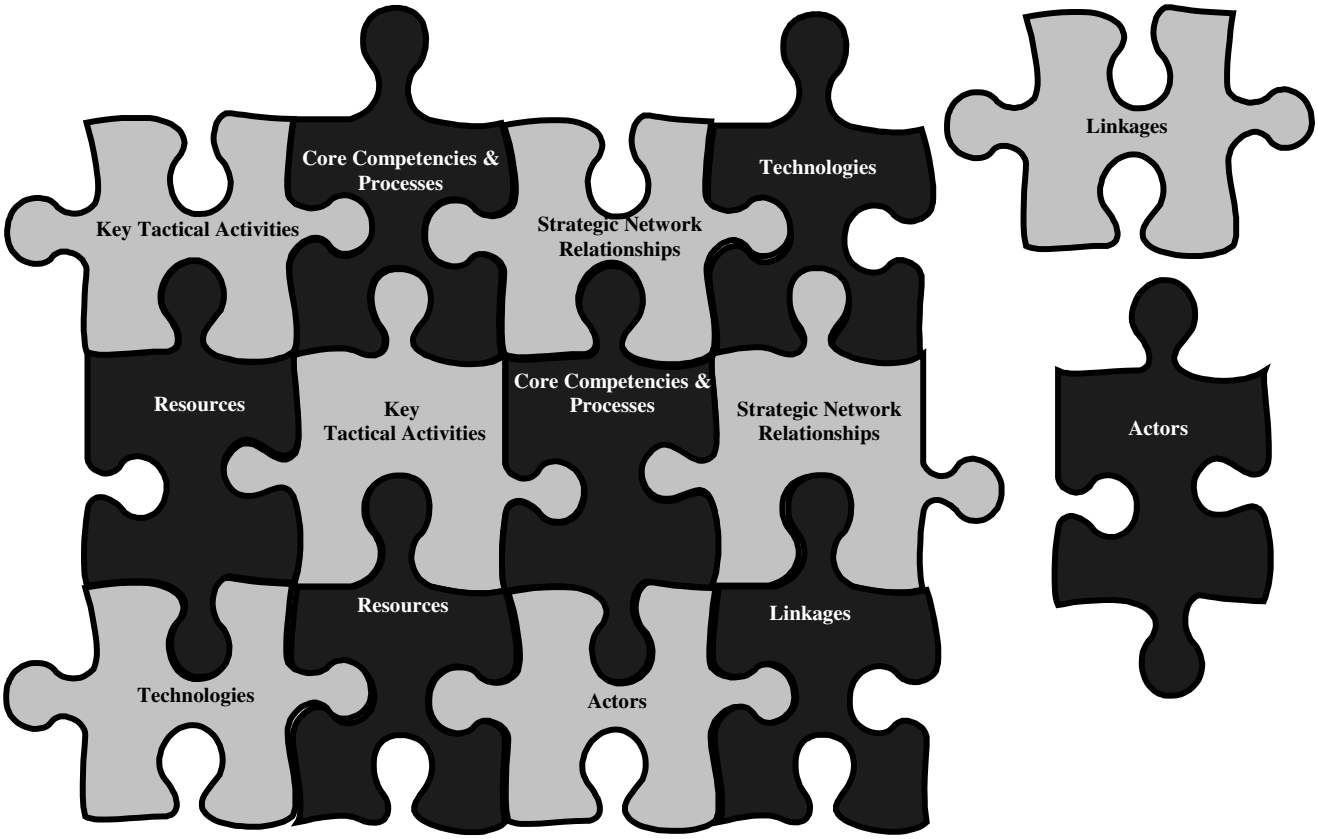


Figure 5: Generic composition of a supply network operations strategy.

to become or remain strategic (this is a fluid situation, requiring constant review). It is clear that increasingly operational effectiveness will rely heavily on supply networks and their key relationships. For many firms, this becomes a strategic issue, as can be witnessed in Volkswagen's Rio de Janeiro plant, which employs 1,000 workers, only 800 of which work for Volkswagen. The rest represent such suppliers as Rockwell, Cummins, and Remon—they actually perform the assembly work on behalf of individual firms.

Resources

Clearly, resources vary by industry and firm, but generically, they can be considered on two levels: the individual resources of the firm (capital equipment, skills, brands, and so on), and the way the resources work together to create competitive advantage. Given the individuality of a resource-based strategy, we have to think of resources as being (a) tangible (physical, technological, financial, etc.), which is important for such producers as Sony and Lexus; (b) intangible (communication and information systems, reputation, culture, brands, and so on)—for investment banks such as Merrill Lynch, reputation is critical; and (c) human (specialized skills and knowledge, communication and interaction, motivation and the like), as is important to Southwest Airlines and the Ritz-Carlton hotel chain.

Technologies

In addition to being a resource used in the general sense (equipment, etc.), technology also includes core technological know-how in product and process innovation across the whole organization and its supply network, as, for example, for Compaq or Microsoft.

The Key Tactical Activities Vital for Supporting a Particular Strategy or Positioning

Certain core tactical activities will be vital to sustaining any particular supply network operations strategy or business positioning. Taco Bell's operations strategy has three main objectives: quality food, good service, and clean environment. However, to deliver these, speed of service at an operational level is equally important and has probably the greatest impact on its revenues.

Actors

Actors range from individuals to groups of companies operating at various levels of control within the network. Daimler-Chrysler has one of the most advanced supply network and outsourcing operations—their workforce and management receive extensive training in these aspects.

Linkages

Linkages are as important as the components. There will be clear linkages and relationships between all the elements in a supply network, and the degree of coordination and control in any particular instance will strongly influence competitive performance. Linkages are usually in the form of shared information and continuous communication. Their importance can be witnessed in the operational approach of a channel assembler, such as IBM,

Dell, and Comapaq. Here, individual components and modules, rather than finished goods, are sent to the distributor, who assembles, tests, and delivers the individual orders.

It seems clear from our research that an organization's operational strategy will evolve over time through dealings with suppliers and customers—indeed many strategic components may, in fact, be introduced at their behest (Lowson, 2001b). Further, the model is dynamic—changing constantly as new building blocks are introduced. However, this inherent complexity provides strength. The complementarity, or cohesion, between the various components (those having both a strong and a weak emphasis) will determine the difference between world-class operations strategies, those that are merely efficient, and those that are suboptimal or dysfunctional.

Despite having common building blocks, these operations strategies will have a unique and individual emphasis that is dictated by a number of factors, such as types of trading partners, supply system configurations, and demand behavior. Here, we can further develop Fisher's (1997) conceptualization of supply chains. He argues that there may be a difference between "physically efficient" and "market responsive" supply chains, based upon their purpose in supplying functional or innovative products and upon the requisite demand patterns. Although Fisher confines his dichotomy to supply chains rather than the full range of operations strategies, he accepts the notion that different supply chains satisfy different demand types. This research allows us to advance a different dimension. First, it seems possible that a firm may use a number of components that are blended into strategic architectures that match the exigencies of the competitive environment. Second, we can also postulate that a supply network operations strategy will have distinct building blocks that provide the necessary and unique emphases to each situation. We can now examine how this strategic customization is undertaken and the important role that supply network relationships play.

Customization of a Supply Network Operations Strategy

The supply network operations strategies employed by a firm in any sector are likely to contain certain components that are combined with particular emphases; it is likely that this act of combination or blending will confer strategic status. If, for example, supply management, time-based competition, and sourcing are of particular strategic importance. An organization will accord operations strategy status to that combination of components. Again, the following points can be made: (a) Some building blocks will be common across a number of commercial sectors dealing with particular products or services. However, there will be others that may be essential to a specific industry that might be peculiar to that setting. (b) Apart from these ingredients of an operations strategy, there will also be other influences on its final shape. The emphasis place upon particular building blocks and their linkages is unique to the operations strategy and provides its sustainable competitive advantage.

These distinct operations strategy "architectures" reinforce Porter's (1996) belief that strategy is mainly a matter of capability development and that activities are the basic units of competitive advantage. An organization, he suggests, will adopt a distinct "strategic positioning"; it will compete on the basis of flexibility, cost, quality, speed, diversity and variety, and so forth. Here, strategic positioning means "performing different activities from rivals or performing similar activities in different ways" (Porter 1996, p 3).

In contrast, he describes "operations effectiveness" as "performing similar activities better than rivals." Clearly, operations effectiveness depends on the correct supply network operations strategy that is a unique blend combining various building blocks. Our premise is that any supply network operations strategy is aimed at performing key operations activities better than rivals, thereby providing support for the overall strategy of a firm, as well as serving as a firm's distinctive competence. Companies can quickly imitate individual activities, but not the way they are combined to form a unique architecture. However, in order to tailor an operations strategy in this way, we must first understand how the building blocks of the strategic architecture can be put together.

What factors or situations dictate whether a particular blend of components is adopted? In other words, do the particular blends vary by demand situation? Developing an optimal operations strategy is clearly a complex challenge. Detailed consideration has to be given to the fusion of building blocks and the emphasis that each can provide in a particular situation. Many firms operating in volatile markets are only now coming to grips with one type of operations strategy; in the future, there will be a number of strategies, each holding a different significance. Perhaps more disturbing, all will be fluid and transitory.

Our research suggests that there are three "drivers" that will shape decisions about the supply network operations strategy and its various building blocks. These are product group demand behavior, supply network behavior, and supply network performance metrics. In Figure 6 we detail a "filtration process" to explain the customization concept.

Figure 6 graphically displays the three main forces shaping a supply network operations strategy; we will deal with each of these in a moment. The driving forces shape the building blocks contained in the strategy (the composition matrix) by influencing both the type of component and the emphasis it is given. In this example, the organization has three strategies, each being customized for a product group and/or customer. These strategies, then, drive future operational management activities in support of a particular product or service. For example, empirical research reported in Lowson (2002, chap. 4) detailed how the Aztec Retail Group, a U.S.-based apparel manufacturer and retailer, embarked on the development of a quick-response operations strategy. However, once work was begun to map the various value streams, it soon became clear that product and customer types were important. The same garment would have different operational needs depending on the type of customer (other retailers, own branches and showrooms, department stores).

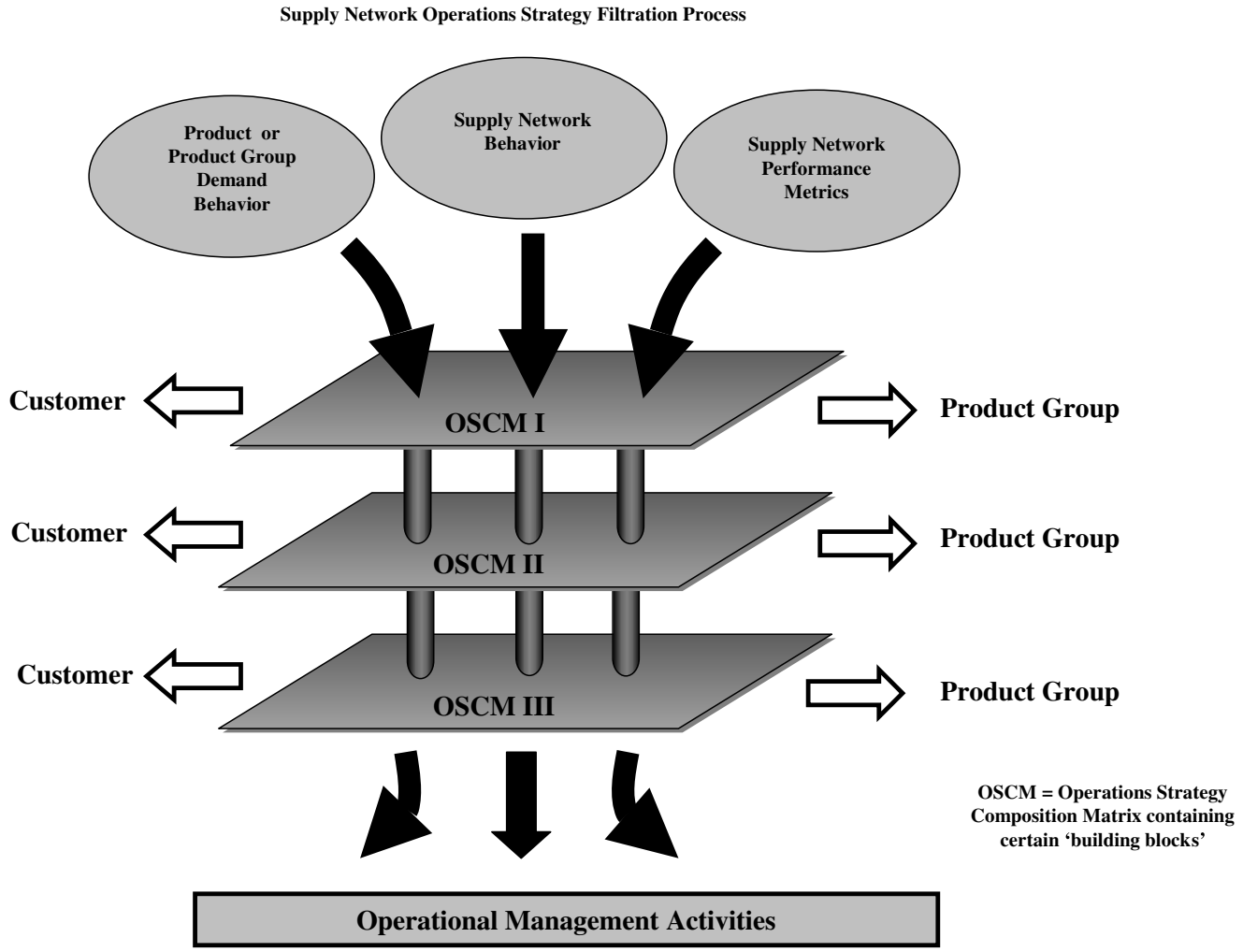


Figure 6: Supply network operations strategy filtration process.

Eventually, the firm adopted three separate operations strategies. We now turn to the driving forces involved in Figure 6.

Product Group Demand Behavior

This category includes product attributes, demand patterns, and customer/consumer behavior.

Product attributes. Product attributes are such things as durability of the product, price levels, shelf life (intrinsic, such as physical characteristics, and extrinsic, such as customer taste changes), product complexity (number and variety of Stock Keeping Units making up a product line), and product line dynamics (characteristics of slow-moving versus rapid-turnover goods).

Demand patterns. Demand periodicity (seasonality, aseasonality, cultural influences), demand uncertainty, and the economics of products lines (unit margins) will all be issues.

Customer/consumer behavior. The different characteristics of the customer (in the case of another organization) and the end consumer (an individual purchaser) place different demands on the supply pipeline and distort the pattern of demand. Different ordering processes, EDI usage, payment methods, logistical needs, promotional requirements, packing and shelving constraints, and product characteristics (sell-by dates, storage needs, transit regulations, maintenance, information content, after-sales support, repeat purchase patterns, etc.) all vary by customer, product group, and individual product. All will change the demand pattern as well as the operations strategy necessary. Add to this the various seasonal effects, cyclical needs, and fashion sways, and the complexity and influence of individual customer behavior begins to be seen.

Supply Network Behavior

Product stream value flows. Product stream value flows are highly complex and dynamic. Any understanding of them requires the appreciation of supply system architecture and the intricacy of that architecture (length of pipeline, stages, firm sizes, product complexity and dynamism, information availability, consumer demands, etc.), the power of the participants (strengths/sizes/financial capabilities of the entities in the pipeline will have a marked impact on the way

it operates), and the integration of the supply system (data/information availability, technology, strategies used, and culture).

Importance of supply network relationships. Supply network relationships are both a driving force shaping the strategy and an important component of it. As was seen earlier, the type of relationship, its purpose, degree of integration and commitment, power balance, strategic importance, strengths and weaknesses, cultural expectations, and so forth, will all exert certain pressures on the shape of the operations strategy.

Vertical integration. The consideration of vertical integration strategies of pipeline members includes direction of any expansion (upstream and downstream), the extent of the process span required (number of supply system activities undertaken by firms), and the balance of the vertically integrated stages (capacity and operating behavior), whether fully or partially balanced.

Supply Network Performance Metrics

Research has now demonstrated that in practice, the building blocks are composed of certain elements (Lowson, 2002). The types of building blocks applied and the degree of significance each receives will be dictated by certain performance factors across the whole supply network. First, each of the individual elements contained within the building blocks can be assessed and weighted in terms of its contribution to competitive advantage (negative, low, medium, and high). Second, the individual building blocks themselves can be evaluated as to performance impact. Here, the measures will be particular to the sector or supply network. Previous studies have, however, included impact on profit, turnover, quality, and customer service levels. The exercise to apply a weighting to these various building blocks and their elements can be conduced at the supply network level; this will include the input of key suppliers and customers as part of a relationship. The supply network operations strategy profiles subsequently created will reflect the differing degree of importance placed on each component with the resulting strategy, thus, becoming customized. Any subsequent supply network operations strategy will incorporate high-scoring elements giving a unique blend and emphases. This exercise can also be conducted by a particular product group and/or customer.

CONCLUSIONS AND FURTHER RESEARCH

This chapter emphasized the strategic nature of supply network relationships. They form an important part of an operations strategy. Such strategies are increasingly adopted to respond to the complex and dynamic trading environment. The chapter began with a review of this domain and a discussion of demand complexity and the need for an organization to offer a flexible response. At that point a model was introduced that demonstrated the evolution of supply systems and gave a future vision: the cluster of value. Using this concept, we considered the response of an organization and the calls for mass customization.

The second part of the chapter addressed supply network relationships as part of a network operations strategy. Here, perhaps for the first time, a genealogy of operations strategies and operational approaches was postulated. This allowed a definition of the supply network operations strategy to be elicited and the role of network relations to be explained. Finally, we described how such supply network operations strategies, and their necessary relationships, can be customized to demand. This process entailed the revolutionary concept that firms will have a number of such strategies and each may be composed of certain building blocks with an emphasis on their interrelationships.

We concluded by describing how the supply network operations strategy and its various relationships could, in practice, be tailored to various demands.

Further research should address a number of key concerns. First, the process by which the various building blocks are chosen must be ascertained. Whether the organization can unilaterally decide for itself or whether it must decide based on powerful suppliers and/or customers is an important point. This has implications for the role and outcome of any relationship forming part of the operations strategy. Second, the performance of the operations strategy itself must be addressed. Its exact impact and the role of relationships on that impact remains far from clear. As the research continues, constructive suggestions from industry and academia are, of course, more than welcome.

GLOSSARY

Building blocks The generic components of supply network operations strategy.

Cluster of value A futuristic vision of the demand and the supply network structure necessary to provide a flexible response.

Filtration process The "drivers" (product or product group demand behavior, network behavior, network performance) that shape the building blocks of a supply network operations strategy.

Flexible response The ability to satisfy demand in all its many forms.

Operational effectiveness Performing similar activities better than rivals.

Strategic positioning Competition on the basis of a unique competence, such as flexibility, cost, and differentiation; the ability to perform activities different from those of rivals, or to perform the same activities in different ways.

Supply network An interconnection of organizations using different processes and activities to add value.

Supply network operations strategy Major decisions about certain medium- to long-term operational capabilities in the supply network and the blending of these into distinct network strategies.

Supply network relationships Key external strategic interconnections and associations formed over time by an organization with its supply networks.

CROSS REFERENCES

See *Electronic Procurement; International Supply Chain Management; Inventory Management; Managing the Flow of Materials Across the Supply Chain; Personalization and Customization Technologies; Strategic Alliances; Supply Chain Management; Supply Chain Management and the Internet; Supply Chain Management Technologies; Value Chain Analysis.*

REFERENCES

Atkinson, J. (1984). Manpower strategies for flexible organizations. *Personnel Management, 2,* 56–78.

BiosGroup. Retrieved June 1, 2002, from http://www.biosgroup.com/solutions/supplynetworks.html

Cap Gemini Ernst and Young. Retrieved June 1, 2002, from http://www.cgey.com

Christopher, M. (1992). *Logistics and supply chain management*. London: Pitman Publishing.

Christopher, M. (1998). *Logistics and supply chain management: Strategies for reducing cost and improving service*. London: Pitman Publishing.

Christopher, M., & Jüttner, U. (2000). Supply chain relationships: Making the transition to closer integration. *International Journal of Logistics: Research and Applications*, *3*(1), 1–23.

Fisher, M. L. (1997). What is the right supply chain for your product? *Harvard Business Review*. March/April, pp. 105–116.

Harland, C. M., Lamming, R. C., & Cousins, P. (1999). Developing the concept of supply strategy. *International Journal of Operations and Production Management*, *19*(7), 650–673.

Lauterborn, R. (1990). New marketing litany: four Ps passé; C-words take over. *Advertising Age*, October, p. 26.

Lee, H. L., & Billington, C. (1992). Managing supply chain inventory: Pitfalls and opportunities. *Sloan Management Review*, *33*(3), 65–74.

Lowson, R. H. (2001a). Retail operations strategies in FMCG supply systems: Analysis and application. *International Journal of Logistics Management, 3*(2), 97–111.

Lowson, R. H. (2001b). Retail sourcing strategies: Are they cost-effective. *International Journal of Logistics*, *4*(3), 271–296.

Lowson, R. H. (2002). *Strategic operations management: The new competitive advantage*. London: Routledge.

Lowson, R. H., King, R., & Hunter, N. (1999). *Quick response: Managing the supply chain to meet consumer demand*. Chichester, UK: John Wiley & Sons.

Porter, M. E. (1996). What is strategy? *Harvard Business Review*, November–December, pp. 61–78.

Shore, B. (2001). Information sharing in global supply chain systems. *Journal of Global Information Technology Management*, *4*(3), 27–50.

Strategic Operations Management Centre. Retrieved June 6, 2002, from http://www.mgt.uea.ac.uk/research/somc

Zipkin, P. (2001). The limits of mass customization. *Sloan Management Review*, *42*(3), 88–88.

T

Taxation Issues

Annette Nellen, *San José State University*

INTRODUCTION

Overview

E-commerce is commerce, which has been taxed for decades, so one might wonder why there is so much discussion today about taxing the Internet and e-commerce. What is unique about e-commerce that has led Congress to debate whether the Internet and online transactions should be taxed, let alone how they should be taxed? Why have there been international conferences and special task forces set up by the Organization for Economic Cooperation and Development (OECD) to enable representatives of many countries to study and discuss taxation of e-commerce? Why are many of the U.S. states finally taking a serious and structured approach to working together to simplify the U.S. sales tax systems due to the advent of e-commerce?

There are a few answers to these questions. First, in many respects, e-commerce is a new business model and existing tax systems were not created with certain aspects of that model in mind. Thus, many existing tax rules and structures do not clearly address e-commerce and Internet transactions. The issues raised exist at all levels—international, national, state, and local. Second, the borderless and global nature of e-commerce will likely require more global cooperation in identifying transactions and developing rational and consistent schemes for taxing some e-commerce transactions. Thus, there is a need for discussions at the international level. Finally, many believe that due to network effects there is a need to keep taxes on Internet transactions low so that the number of users grows and it becomes a better and more valuable network. That is, the Internet is a more useful tool if there are more and more people on it with whom you can communicate and transact business.

This chapter explores the specific reasons why e-commerce and the Internet raise tax issues, reviews tax fundamentals and legal constraints that must be considered in resolving these issues, and explains the basics of the key issues that exist at each level of government for different types of taxes. In addition, significant developments in resolving the issues are explained.

The primary reasons why e-commerce and the Internet raise tax issues are discussed next, along with the reasons why the tax issues need resolution and what fundamental tax and constitutional principles must be considered in resolving them.

Why E-commerce and the Internet Raise Tax Issues

Location

Existing tax systems tend to determine tax consequences based on the physical location of the taxpayer. The e-commerce model enables businesses to operate with very few physical locations, and thus, fewer taxing points. An online vendor can easily sell to customers throughout the world from a single physical location. The e-commerce business model also involves more customized inventories so storage needs (and thus the need for many physical locations) are reduced. Online sales reduce the need for firms to have sales offices scattered throughout a sales region. Also, the model involves less vertical integration and more outsourcing—again, resulting in fewer physical locations necessary for a vendor to operate. Also, some business assets, such as servers, are not necessarily tied

to a single physical location, but can easily be relocated without any interruption to business operations. That is, the location of the server is not relevant for business purposes and thus, may not be a logical taxing point.

Location factors primarily raise tax issues at the international, state, and local levels. For example, the U.S. Supreme Court has ruled that a state may only require a vendor to collect sales and use tax if the vendor has a physical presence in the state (*Quill*, 1992). With fewer physical locations, states will need to work either to have the physical presence standard changed by Congress or to get resident consumers to voluntarily remit the use tax that is owed when a vendor was not required to collect sales tax (the use tax is a complement to the sales tax in all states with a sales tax). Collecting use tax from buyers is difficult due to the greater number of taxpayers (relative to collecting the tax from vendors) and difficulties of getting consumers to understand the tax base. Also, outsourcing of various services raises issues as to the nature of the relationship between the company and the supplier in order to determine whether the supplier is the company's agent, thereby creating a taxable presence (nexus) for the company in the state.

Another potential issue tied to location is that the workforce of an Internet-based company may be scattered throughout a state or country. This can raise issues as to whether the presence of the employee in a particular state creates tax obligations for the employer in that state. Also, cities may find that employers owe business license taxes due to the presence of an employee in the city or if the worker is not an employee, that the worker owes business license taxes. Cities will likely have difficulty finding these workers and businesses due to the number of people involved and their lack of a storefront (many may be working out of their home).

Nature of Products

E-commerce allows some types of products, such as newspapers and music CDs, to be delivered in digitized (intangible) form, rather than tangible form. Digitized products raise issues at the state level as to whether sales tax applies and in which state income is generated for state income tax purposes. Public Law 86-272, enacted in 1959, prohibits a state from taxing a foreign corporation's net income derived from activities within the state if those activities consist merely of solicitation of orders for the sale of *tangible* personal property approved, filled, and shipped from outside the state. The nature of products can also raise income tax issues regarding the type of revenue generated and how it is to be reported (sale of goods, sale of intangibles, or services), as well as how digitized products are reported under the tax accounting rules for inventory.

Nature of Transactions

The Internet has facilitated innovation in the sale and purchase of goods and services. For example, individuals can offer items to a worldwide group of potential buyers via auction sites, such as eBay. The Internet can also be used to easily link business buyers and sellers through exchange Web sites where buyers post what they have to sell and sellers match up with them, or vice versa. Such sites can almost operate without human intervention for the matching and payment functions. In addition, the Internet has increased the use of bartering, most notably with respect to exchange of Web banners that serve as advertisements. These new techniques raise various tax issues. For income tax purposes, issues include whether an exchange intermediary or broker should be accounting for inventory, what amount of information reporting should be required for low-value bartering transactions, and how such transactions should be valued. At the international level, the source of the income generated (which country) might be uncertain. For example, should income generated from a product exchange Web site be attributed to the country where the servers are located or where employees or customers are located or some combination of these locations? At the state and local level, issues exist as to when individuals have sold enough goods to be required to become sales tax collectors and how to enforce such rules. Another issue raised by changes or elimination of intermediaries is that some intermediaries collect excise tax, such as sellers of fishing equipment. When buyers interact directly with a foreign manufacturer, rather than a domestic retailer, the excise tax may go uncollected.

The Internet also allows for paperless transactions and the potential for the use of electronic cash. This raises administrative concerns for the Internal Revenue Service and other tax agencies as to whether transactions were properly reported, whether an audit trail exists, and whether new reporting rules are needed. The concerns of tax agencies include the reality that their technology may not be as up-to-date as that used by taxpayers, digital transactions may be difficult to identify or track, and the ease of operating from various locations may make it easier for tax evasion to proliferate.

Why the Issues Need Resolution

A report on e-commerce from the European Union (1997, paragraphs 56, 58) posited that to enable e-commerce to develop, tax systems needed to provide legal certainty, tax neutrality, and safeguards for government revenues. These three factors have also been identified by the U.S. and other countries as key e-commerce taxation concerns. These three factors or goals help to identify the nature of the work that needs to be done to provide e-commerce businesses, consumers, and governments the certainty needed to operate effectively and have confidence in the e-commerce business model. These three factors are explained next along with some data to illustrate the underlying concerns.

Legal Certainty

If businesses have uncertainties as to how e-commerce and Internet transactions are taxed and in which jurisdiction, the highest use of e-commerce as a business model will not be achieved. Also, if consumers are unsure as to whether something is subject to sales tax or if they should self-assess a use tax, they may be reluctant to buy goods and services online. Thus, there is a need to clarify and perhaps modify existing tax rules and systems so they better address Internet and e-commerce transactions.

Neutrality

The neutrality argument is often described as needing to "level the playing field." For example, if consumers can purchase books from an online bookseller that has no physical presence in the state where the consumer resides, no sales tax need be collected. In contrast, if the consumer walks into a bookstore and purchases a book, sales tax will be collected. Thus, neutrality does not exist because the tax result may influence the individual's decision on how to buy the book (although this might be mitigated by the shipping and handling costs typically charged by Internet vendors). Many "Main Street" vendors would like to see Congress enact legislation that would allow states to collect sales and use taxes from nonpresent (remote) vendors in order to "level the playing field." Due to the complexities of the sales tax rules, which vary somewhat among the over 6,000 jurisdictions in the United States that impose this tax, Congress is unlikely to impose additional collection burdens on vendors until some simplification is achieved (and even then, of course, the nature of any congressional action is uncertain at this point).

For companies that have collection responsibilities due to a physical presence in most, if not all, of the 46 states imposing a sales tax, compliance costs can be very high. A multistate company will need to hire several individuals to satisfy its compliance obligations, acquire software to assist in the process (which requires at least monthly updates), and to keep up with changes in the law in each state in which it does business. Large companies with many physical locations must budget the resources to comply with all of these tax systems. However, as noted earlier, the Internet allows new and small vendors to operate in many states without needing any capital for physical (bricks and mortar) expansion. Such vendors would likely be put out of business or not seek the market potential presented by the Internet if faced with obligations to collect sales and use taxes from customers in 46 states and multiple local jurisdictions. While software solutions are available, the cost, along with the need for personnel to operate the software, means that this is not a feasible solution for most small businesses.

The task of reducing the existing complexities for multistate vendors are challenging because in some states, such as Colorado, cities define the tax base, which could (and does) differ from the state tax base. Also, local tax rates are not always bound by zip codes. Instead, some zip codes can include more than one local tax rate. In addition, definitions of items (such as handling costs) differs from state to state, as do exemptions from the tax. In addition, tax forms and due dates, as well as audit procedures, vary from state to state.

The costs of complying with the tax rules of multiple taxing jurisdictions can be significant in terms of labor costs, training, computer systems, need for continual updates due to changes in laws and regulations, audit costs, credit card fees, and risk of error. A 1998 study by the State of Washington on sales tax compliance costs determined that compliance costs as a percent of total state and local sales tax collections were as follows (Washington State, Department of Revenue, 1998):

Small business 6.47%	(gross sales between $150,000 and $400,000)
Medium business 3.35%	(gross sales between $400,000 and $1,500,000)
Large business 0.97%	(gross sales over $1,500,000)
Total cost weighted by number	4.23%
Total cost weighted by dollars	1.42%

The significance of the administrative costs to vendors means that if collection obligations are to be expanded (such as by requiring remote vendors to collect sales tax), vendors will continue to demand simplification, as well as compensation for the costs of collecting the government's revenue. The greater costs for small businesses (relative to their sales) support consideration of collection exemptions for small businesses. Of course, such an exemption would violate the neutrality principle unless the customers of small businesses were required to remit the tax (which raises a complexity issue). This would be an example of the need to weigh the benefits of simplification and neutrality, since a government is unlikely to achieve the ideal tax system.

Other taxes being looked at in the discussions of e-commerce and Internet are telecommunication taxes imposed by federal, state, and local governments. The compliance burden can be fairly high for these types of taxes. A study by the Council on State Taxation (COST, 2002) found that the national average effective tax rate for transaction taxes on telecommunications services was 16.90% compared to 6% for sales of goods. In addition, filing obligations in the state and local jurisdictions totaled 66,918 returns for telecommunications taxes compared to 8,284 returns for general businesses.

The COST study also found that telecommunications companies have 391 types of taxes imposed upon them by state and local jurisdictions, compared to 118 for general businesses. In addition, only 16 states have a sales tax exemption for communications equipment while 37 states have an exemption or reduced rate for manufacturing equipment. These results point to a neutrality issue to the extent that regulated telecommunications companies may be subject to more taxes than nonregulated companies that may provide similar services (for example, an Internet service provider (ISP) providing Internet telephony that is not subject to telecom excise taxes).

Revenue Protection

A final reason why e-commerce taxation issues require resolution is that there is a potential for revenue loss at each level of government. For example, because the e-commerce business model enables businesses to operate with few physical locations, yet reach customers throughout the world, states will find that there are more nonpresent vendors not required to collect sales tax. Although consumers purchasing taxable items from nonpresent vendors should self-assess use tax and remit it to the state, use tax compliance and enforcement is very low and it is difficult to collect the tax from consumers. Several states do attempt to collect the use tax by having

individuals report it on their state personal income tax form, but compliance is still below 100%.

Various studies have been done privately and by governments to identify the costs to government of the inability to collect sales and use tax on many e-commerce sales. The results of a few of these studies are explained next.

A study by Ernst & Young (Cline & Neubig, 1999) concluded that 63% of business-to-consumer online sales were nontaxable (such as airline tickets, gambling, and interactive games). Of the remaining 37% of business-to-consumer sales, sales tax was paid on 4% (4% of the 100% of business-to-consumer sales), and 20% was a substitute for other remote sales for which no tax was collected, leaving 13% of total business–consumer sales untaxed. The study applied an average state and local sales tax rate of 6.5% to determine that the estimated sales tax loss was $170 million for 1998, representing one-tenth of 1% of total state and local sales tax collections.

A General Accounting Office (GAO) report (2000) estimated that the state and local sales and use tax losses for all Internet sales for 2000 was between $0.3 and $3.8 billion (about 2% of projected sales tax revenue). This included both business-to-business and business-to-consumer Internet sales. The projected loss for 2003 was between $1.0 and $12.4 billion (5% of projected sales tax revenue). The difference between the high and low figures is due to varying assumptions concerning business-to-business compliance rates and the estimated amount of e-commerce sales.

Although this revenue loss is low in the early stages of the e-commerce business model, governments are concerned that the likely growth of e-commerce will cause the revenue losses to become more significant. Also, many states and cities face restrictions in changing their tax structure (such as constitutional prohibitions against some types of taxes or the need for a majority vote of taxpayers to raise a tax). Thus, many are eager to see issues resolved within existing tax systems and to see resolution of e-commerce and Internet tax issues sooner, rather than later when the potential revenue losses increase.

Considerations in Addressing the Issues

Resolution of the tax issues raised by the Internet and e-commerce needs to consider the principles of good tax policy, restraints imposed by the U.S. Constitution, and the global business environment in which e-commerce operates.

Tax Policy

Any discussion of tax reform should consider the tax policies that generally underlie every type of tax structure to some degree. The American Institute of CPAs (AICPA) has identified 10 principles of good tax policy (AICPA, 2001), partially based on Adam Smith's maxims of good tax policy. Principles most relevant to the e-commerce taxation debate include the following:

1. *Certainty*—Taxpayers should have a high level of confidence that they are calculating their tax liability correctly. When e-commerce vendors are unsure as to whether they have a taxable presence in a state or where to source their operations, certainty does not exist.
2. *Convenience of payment*—A tax should be due at a time or in a manner most likely to be convenient for the taxpayer. For example, a tax upon the purchase of goods should be assessed at the time of purchase when the person still has the choice as to whether to buy the goods and pay the tax. Today, use of technology to collect tax should be considered for transactions that already depend upon technology for their execution.
3. *Simplicity*—A tax system should be simple so that taxpayers can understand the impact of taxes on their transactions. When considering state and local taxes, uniformity should also be considered to reduce or eliminate a multistate taxpayer's complexity of dealing with multiple sets of tax rules and filing procedures.
4. *Neutrality*—The tax law should not impact economic decisions. That is, taxpayers should not be unduly encouraged or discouraged from engaging in certain activities or taking certain courses of action primarily due to the effect of the tax law on the activity or action. For example, the tax consequence of purchasing a digitized book should be the same as for purchasing a physical book, assuming users view them as equivalents.
5. *Economic growth and efficiency*—The tax should not discourage or hinder capital formation, employment, and investment. Tax rules should not discourage the growth of e-commerce if governments, businesses, and consumers view this vehicle as something to be encouraged.
6. *Appropriate government revenues*—The tax system should enable the government to determine how much tax revenue will likely be collected and when. The tax system(s) should be capable of raising the necessary government revenue. E-commerce has and will continue to affect government tax collections. State and local governments will experience drops in sales and use tax collection as residents can more easily make online purchases from remote (nonpresent) vendors and are unable to collect use tax from the buyers. States will see some companies eliminate physical location (taxing points) as sales offices and warehouses are closed in response to increased online sales and custom orders. Of course, as companies consolidate taxing points in fewer states, those states should see an increase in revenues. A similar effect will occur worldwide as companies consolidate their taxing points (permanent establishments).

It is unlikely that all of the principles of good tax policy can be achieved equally under any tax system. For example, in some instances, it might be advisable to have simplicity override neutrality, such as by exempting small businesses from a certain system even though doing so would undermine neutrality of the system.

Constitutional Constraints

The U.S. Constitution imposes some constraints on state and local government taxation. The due process clause of the 14th amendment requires fairness by a government in imposing its laws upon taxpayers. Due process requires some minimum connection between the taxing state and the taxpayer such that it would be fair to impose the tax rules upon the taxpayer. The Commerce Clause

gives Congress the power to regulate commerce among the states. In determining whether the due process clause is violated, courts will look at the effect of the suspect tax rule on the particular taxpayer involved. In contrast, to determine whether the Commerce Clause is violated, the courts will look at the effect of the particular tax rule on the national economy—does it impede commerce among the states, such as where a state is taxing activity that occurs outside of its borders?

A state or local tax that creates a substantial risk of international multiple taxation or preventing the federal government from speaking with one voice in the regulation of commercial relations with foreign governments may violate the Commerce Clause (discussed later), Export Clause (no tax or duty may be imposed on articles exported from any state), and/or Supremacy Clause (federal law is the supreme law of the land).

Global Business Environment

Because the e-commerce business model is a global one for businesses of all sizes, any country or state looking at resolving e-commerce taxation issues should be cognizant of what similar proposals and actions are being considered by other countries and at the international level, such as by the OECD. In the e-commerce taxation studies that several countries engaged in during the late 1990s, it was recognized that countries will need to work together to deal with tax issues so as to avoid multiple taxation and undue competition, as well as to update and coordinate treaty provisions, enforcement efforts, and the legal basis for taxing multinational transactions.

An OECD report (1998) noted that broad taxation principles should apply to e-commerce. With respect to consumption taxes, two of the conclusions reached are relevant not only to the U.S. Government, but also to state and local governments. First, the OECD concluded that consumption taxes should be based on the location where consumption takes place. Second, digitized products should not be treated as a supply of goods (but as a service for consumption tax purposes).

TAX ISSUES RAISED BY E-COMMERCE AND THE INTERNET

The general reasons why the Internet and e-commerce raise tax issues, such as location and digitized products, were explained earlier. This section covers the tax rules that raise issues. The issues are grouped together by the nature of the issue along with the specifics of the issues at the international, national, and/or subnational levels. The subsequent section explains some of the activities underway to resolve these issues.

Authority to Tax—Nexus and Permanent Establishment

In any tax system, it is important to know which jurisdiction has authority to impose a tax on a business or individual. The federal government and most states have provided some guidance with respect to both sales and income taxes. The terminology commonly used is nexus, which involves knowing in which state it would be fair to impose tax collection or payment obligations on a taxpayer. At the international level, countries use the concept of permanent establishment to determine when a business should be taxed within a particular jurisdiction because the taxpayer has a place of business there.

Nexus

For sales and use tax purposes, the U.S. Supreme Court has held that a business must have a physical presence in a state before the state may impose sales tax collection requirements on a vendor. The ruling (*Quill*, 1992), involved Quill Corporation. Quill sold office products to customers in North Dakota and did not have a sales office, warehouse, or personnel in North Dakota. Quill conducted business by sending product catalogs to companies in the state. In determining whether it was permissible for North Dakota to collect sales tax from Quill, the Court noted that for due process purposes, it would be more appropriate to not focus on physical presence, but to instead look at whether the company's contacts with the state make it reasonable for the state to require the company to collect use tax. In *Quill*, the Court stated that if an out-of-state business purposefully avails itself of the benefits of an economic market in the state, it need not have a physical presence in the state to be subject to tax collection requirements in the state.

With respect to the Commerce Clause, the Court stated that North Dakota's enforcement of the tax against Quill was an unconstitutional burden on interstate commerce. However, the Court pointed out that because the Constitution gives Congress the right to regulate interstate commerce, Congress could provide a mechanism to allow states to collect sales and use tax from an interstate mail-order business that was not physically present in the state, without violating the Commerce Clause.

Issues still exist after the *Quill* decision because it is not clear how much physical presence is needed to enable a state to impose sales tax collection responsibilities on vendors. Quill actually had some diskettes in the state, but the Court did not view that as sufficient physical presence to allow the state to collect tax from Quill. Issues include whether software is considered a physical presence (some states have taken the position that it is tangible), whether leased phone lines and servers create nexus, and how often employees must be in a state before the employer is considered to have nexus.

With respect to state income tax, Public Law 86-272 (discussed earlier) provides guidance on when a state may impose income taxes on a business with customers within the state. Because this rule only applies to sales of tangible property though, it is out-of-date today because the transfer of services and intangible items is a significant business activity. Where a transaction does not fall under the protection of this public law, the general guidance on nexus is to be applied.

Permanent Establishment

In 2000, the OECD addressed the concept of permanent establishment (PE) in an e-commerce environment by drafting additional commentary to become part of its model tax treaty (OECD, 2000). Tax treaties between countries serve to identify whether the resident or source country has the right to tax income of a multinational business in order to avoid double taxation. Most tax treaties today

incorporate the principles and terms of the OECD model tax treaty.

The new commentary on PE and e-commerce provides that a Web site does not create a PE. It also clarifies that human intervention is not required to have a PE in a country, thus opening the door to the possibility of equipment alone (such as servers) creating a PE. However, the facts and circumstances would have to be such that the equipment performed functions that were more than preparatory or auxiliary to the generation of income. If a server is merely hosting a Web site or displaying a product catalog, it is unlikely to be a PE. However, if a server is enabling the processing of orders and payments and perhaps delivery of product, it is likely to be a PE. Because the OECD treaty is a model, rather than a mandate, it remains to be seen whether all countries will adopt the same perspective as presented in the new commentary. Also, the OECD continues to explore whether additional guidance is needed to determine how much income is attributed to a PE of an online vendor.

What to Tax—Tax Base and Characterization of Income

To calculate any tax liability, one needs to know both the tax rate and tax base. The well-known, but poorly understood, "Internet tax moratorium" also has some bearing on a state tax base, as described below. Finally, one often needs to know the particular type of income involved (sales of goods, services, or royalties) to determine tax consequences.

Sales Tax Base

One of the most complicating factors about the sales tax for multistate businesses is that not only do sales tax rates vary from state to state and even within different cities in a state, but the tax base varies from state to state (and perhaps even between a city and state). For example, the number and types of services taxed by the states vary tremendously. Hawaii, New Mexico, and South Dakota tax over 140 types of services, while 7 states tax fewer than 20 types of services (states without a sales tax have been omitted from this count). Although most states do not tax custom software, 15 states and the District of Columbia tax it. Some states tax computer-related services (FTA, 1997).

Many states exempt food, even more exempt prescription drugs, and a few also exempt nonprescription drugs. A few states, such as Illinois, tax these items, but at a lower rate than for other taxable items. In California, clothing is subject to tax, although it is exempt in Minnesota, and in Massachusetts, clothing is only subject to sales tax to the extent its cost exceeds $175 (information obtained from state tax agency Web sites: California, http://www.boe.ca.gov/pdf/pub61.pdf; Massachusetts, http://www.mass.gov/dor/publ/pdfs/sls_use.pdf; and Minnesota, http://www.taxes.state.mn.us/salestax/factshts/salestax.html). In addition, the definition of clothing may vary from state to state. For example, it may or may not include recreational clothing, such as sports gloves and ski boots.

The complexity of the rules and cost of compliance can cause new online vendors to limit the number of states in which they have a physical presence and, thus, a sales tax collection obligation. Efforts have been underway to streamline the sales tax rules, including definitions for the tax base, but it is uncertain how many states will participate in the reforms. Also, even with streamlining, tax bases will continue to differ although the definitions of what is and is not taxable would be the same from state to state.

The "Internet Tax Moratorium"

With respect to the Internet, Congress has exercised its authority under the Commerce Clause to impose a moratorium on state and local taxes on Internet access, unless such tax was generally imposed and actually enforced before October 1, 1998 (about 10 states had such taxes in place). The moratorium also applies to multiple or discriminatory taxes on e-commerce. This moratorium was created by the Internet Tax Freedom Act (Public Law 105-277) in 1998, and was extended to November 1, 2003, by the Internet Tax Nondiscrimination Act (Public Law 107-75).

The moratorium was enacted to ensure fair and administrable rules so that growth of the Internet would not be impeded. Also, proponents of the Act wanted to avoid multiple taxation of transactions among the states and to allow time for thoughtful consideration on how the Internet and online transactions should be taxed. Opponents of the Act noted that it imposes a federal restriction on state and local governments and could reduce state and local tax revenues. A report by the Center on Budget and Policy Priorities (Mazerov & Lav, 1998) noted that the Act allows for unfair competition by online vendors at the expense of Main Street vendors, benefits wealthier consumers who tend to use the Internet, and is not needed as there was no evidence that state and local taxes were impeding the growth of the Internet and e-commerce.

The moratorium has created a fair amount of confusion, even leading some consumers and businesses to believe that the Internet and e-commerce are tax-free. However, this is not the case. The moratorium primarily prohibits state and local governments (unless they fall within the grandfather provision) from taxing Internet access services. The Act specifically preserves state and local taxing authority that is otherwise permissible. Thus, purchases made via the Internet, if otherwise taxable, are still taxable when purchased "on the Net." For example, if a California consumer purchases books from a bookseller that only has a physical presence in New York, no sales tax is collected (because the vendor has no physical presence in California). However, the buyer is required by California law to self-assess a use tax (equivalent to the sales tax). It is the *Quill* decision (discussed earlier), not the moratorium, that absolves the New York vendor of the responsibility to collect sales tax from its California customers.

In addition, even with the "Internet tax moratorium," Internet access providers and other Internet businesses are subject to federal and state income taxes, business license taxes, property taxes, and payroll taxes, just like other businesses. Purchases of airline tickets via the Internet are included in the tax base for federal excise taxes

and phone bills still include excise taxes even if part of the phone usage was to access the Internet.

Finally, the Act included a declaration that no new federal taxes similar to the state and local taxes covered by the Act should be enacted with respect to the Internet and Internet access during the moratorium.

Characterization of Income

At the international level, a Technical Advisory Group (TAG) set up by the OECD has issued guidance on classifying the type of revenue for purposes of being adopted as commentary to the OECD model tax treaty. The report (OECD, 2001) points out that the issue of determining whether payments are for business profits or royalties is not a new issue. Instead, this has become a more commonplace issue with the advent of e-commerce where it is easy to transfer digital items.

The main issue is to distinguish when the making of a copy of a digital item represents royalties for use of a copyright rather than normal business profits. Because the act of copying involves use of the copyright, some countries have taken the position that the payment is a royalty. The TAG recommends that where a payment for a digital item is made so that the buyer can use and enjoy the item, the use of the copyright is an incidental rather than primary purpose of the payment and should not be the focal point to characterize the payment.

Where to Tax—Sourcing

International Perspective

At the international level income sourcing rules dictate whether income is taxed in the country of residence or source. Tax treaties typically provide that the source country has preference over the resident country in taxing business income if there is a PE in the source country (rules may vary for other types of income). Several countries have noted in reports on e-commerce tax issues that the sourcing rules may need to be revisited for some types of e-commerce transactions. Coordination of sourcing rules is needed to avoid double taxation of income (such as where income is taxed both in the country of residence and source).

In 1996, the U.S. Treasury issued a report that included the following suggestion:

> The growth of new communications technologies and electronic commerce will likely require that principles of residence-based taxation assume even greater importance. In the world of cyberspace, it is often difficult, if not impossible, to apply traditional source concepts to link an item of income with a specific geographical location. Therefore, source based taxation could lose its rationale and be rendered obsolete by electronic commerce. By contrast, almost all taxpayers are resident somewhere. (U.S. Treasury, 1996, ¶7.1.5)

Most countries did not embrace this perspective and the Treasury Department later backed down from the statement. One concern was that because so many online vendors are located in the United States, the United States would gain revenue from a residence-based sourcing system for e-commerce.

Countries may see a loss of revenue because the e-commerce business model does not require a physical presence (property or employees) in a location in order to serve customers at that location. For example, a multinational company may be able to eliminate a PE in a country by closing a sales office and having customers purchase online while the taxpayer's equipment is located in a low-tax country.

Another international "where" issue exists for some multinational businesses when more than one country treats the business as a resident. Under the OECD model tax treaty, a business is deemed to be a resident of the country where its place of effective management resides. This is the location where key management and commercial decisions are made—typically the location of the key executives. A business may have more than one place of management, but under the model treaty, it will have only one place of effective management based on the facts and circumstances.

The e-commerce business model challenges the determination of the place of effective management because key executives may not all be in the same location and may not be in the location where key business operations are conducted.

Discussion of the issues of sourcing international business income will continue. Global cooperation in establishing rules will prevent double taxation and, perhaps, undue competition among countries. The OECD and its TAGs continue to work on identifying the issues and possible solutions. The draft and final reports are available at the OECD web site (OECD, n.d.).

State Income Tax Perspective

Sourcing rules are also relevant at the state level to determine which state (or states) may assess income tax on a multistate business. These rules are typically different for income from sale of goods than for services and intangibles.

Today, for income tax purposes, most states source sales of tangible products to the destination state (in most states, if the seller is not subject to tax in that state, the sale is "thrown back" to the state of origin). Sales of intangibles are typically sourced to the state where the greatest income-producing activity occurs. However, a few states, such as Minnesota, source revenue from services and intangibles to the state where the service is received or the intangible is used by the purchaser.

While almost all of the discussions concerning state taxation of e-commerce focus on sales and use taxes, there are also several state income tax issues, particularly with respect to sourcing of income from e-commerce vendors. With respect to intangibles, not all states determine the location of the greatest income-producing activity in the same manner. Some states look at where the greatest cost of performance occurs, while other states prorate income based on the portion of the costs of performance that occur in the state. In addition, states do not consistently measure the costs of performance in that some only consider direct costs while others also include indirect costs.

Also, not all states treat items transferred online as tangible or intangible in determining which sourcing rules apply. Finally, under the e-commerce model where one piece of equipment may be handling most of the sales, processing, and delivery functions, it is not clear whether such costs should end up being the determinant of where the greatest costs of performance occur or whether employee wages should be a more significant determinant. Discussions here need to consider that equipment is generally more mobile than personnel and the premise for determining what factors indicate that services and intangibles revenue was generated in a particular location (or locations).

Consumption Tax Perspective

Sourcing rules are also relevant for consumption taxes to determine whether the transaction should be taxed at the vendor's location (origin basis) or customer's location (destination basis). In the United States, the states use the destination approach to determine where sales tax should be assessed. Other countries tend to do the same. However, some have suggested that the origin approach may be more suitable for online transactions, particularly for transfers of digitized products and services. Where an item is transferred electronically, it is far easier to determine and verify the vendor's location than the customer's location. Thus, for compliance purposes, the origin approach seems to be simpler. However, others raise the issue that the origin approach may result in a "race to the bottom" where states reduce their sales tax rate to attract businesses and in the process, hurt state tax revenue collections. The origin versus destination issue will continue to be discussed both at the state and international levels.

Additional Tax Issues

The issues described previously are those most frequently discussed in e-commerce taxation debates. However, there are a variety of other issues at all levels of government. A few significant issues are briefly described next.

Web Site Development Costs

For income tax purposes, the law is not clear as to when Web site development costs must be capitalized rather than expensed. Also, if such costs must be capitalized, what is the depreciable life? Although guidance exists on the treatment of software development costs, it is not clear whether all Web site development activities constitute software development (a term that is not defined for federal income tax purposes).

Domain Names

There have been reported stories in the press of individuals selling domain names for significant amounts of money. Often, the individuals have-registered many domain names in the hopes that some business will approach them to purchase a name at a premium. Tax issues can arise for both the seller and buyer. The seller will need to review the definition of a capital asset to determine whether sale of the name produces capital gain income or ordinary income. The buyer will need to determine what the depreciable life is for the name. Existing guidance on depreciation of intangibles (Internal Revenue Code §197) created in 1993 does not clearly provide an answer.

Who to Tax

The issue of who to tax generally only arises for consumption taxes. Although these tax rules are usually very clear as to the answer, that answer may not be viewed as practical by the taxing jurisdiction. For example, although consumers who are not charged sales tax on an otherwise taxable online purchase where the vendor does not have a physical presence in the state still owe a use tax, such a tax is difficult to collect. States would much prefer to be able to have vendors collect the tax (fewer taxpayers to deal with) than try to collect the use tax from consumers. The record-keeping burden and need to understand the tax rules make it difficult to collect the use tax from consumers.

The European Union, in particular, has been concerned about the collection of value-added tax (VAT) on digitized products and services sold by non-EU vendors to individual consumers in the EU. The issue is similar to that in the United States of states being unable to collect use tax from nonpresent vendors, and the desire of EU vendors to "level the playing field" between online vendors and Main Street vendors.

In 2002, the EU announced final rules to address the concern. Under these rules, a non-EU vendor of digitized items and services (such as software, games, and broadcasts) must register in at least one of the 15 EU Member States and collect VAT at the rate specified in the place where the nonbusiness customer resides. Non-EU vendors are not required to collect VAT from business customers because such customers are to self-assess the VAT. The U.S. Treasury has expressed concern over the EU rule, noting that it creates a compliance burden and leads to some discrimination in that in some EU States digitized products are not taxed at the same VAT rate as the equivalent tangible product. In addition, the Treasury notes that EU vendors are allowed to assess VAT based on the rate in the vendor's country, which will lead to increased competition for non-EU vendors, who must charge a higher rate in instances where the customer resides in a higher rate jurisdiction. Finally, the Treasury notes that it would be best for countries to act together to resolve these issues, such as through the OECD, rather than resolve them separately (U.S. Treasury, 2002).

Administrative Considerations

Tax authorities have identified a wide range of administrative concerns that potentially exist in the e-commerce environment. These include issues of tracking transactions where electronic money is used, existence of an audit trail, viability of a digital signature system for tax filings, reliability and verification of digital records, identification of cross-border digital transactions, information reporting when a typical reporting intermediary no longer exists, reporting of small value bartering transactions (such as for Web banner ads), and increased mobility of taxpayers.

The OECD and many countries continue to study many of these taxation issues. Resolution of many of these issues will take some time and probably global cooperation.

ACTIONS TAKEN TO RESOLVE INTERNET AND E-COMMERCE TAX ISSUES

As described so far in this chapter, there are a number of Internet and e-commerce tax issues in need of resolution at all levels of government. The issuance of the Treasury study in 1996 followed by issuance of similar reports by other industrialized countries, as well as work done by a federal commission in 1999–2000, have led to helpful discussion of the issues and possible solutions. Notable activities are briefly summarized below.

Federal Advisory Commission on E-commerce

In addition to imposing a moratorium, the Internet Tax Freedom Act (described earlier) created a 19-member Advisory Commission on Electronic Commerce (ACEC). The members included 8 representatives of state and local governments (including one from a state with no sales tax and one from a state with no income tax), and 8 representatives from the e-commerce industry (including small business), telecommunications carriers, local retail businesses, and consumer groups. The remaining 3 members were the Secretary of Commerce, the Secretary of the Treasury, and the United States Trade Representative. The ACEC was to conduct a thorough study of all levels of tax with respect to e-commerce and report to Congress in April 2000. The ACEC report could contain legislative recommendations if they were tax and technologically neutral and were approved by at least two-thirds of the Commissioners.

In April 2000, the ACEC issued a report describing proposals of the majority. The ACEC was unable to reach the required 2/3 vote to come up with recommendations on most matters. The report did not include the views of the minority. While the ACEC failed to provide legislative recommendations, the discussions and testimony that arose from its existence and the issuance of majority and minority reports did heighten awareness of the issues and possible solutions. Common themes in both reports included the need to simplify sales tax and telecommunication tax systems and to not tax Internet access fees. The majority and minority reports, along with the testimony presented to the Commission, provide a rich source of information on the nature and range of e-commerce taxation issues, as well as possibilities for resolving many of these issues (ACEC, n.d.).

Streamlined Sales Tax Project (SSTP)

The SSTP involves a group of representatives from over 35 states and the District of Columbia who are working together to create a Model Act and Agreement for a uniform and simplified sales and use tax system (SSTP, n.d.). Language was approved by the participating states in December 2000. While additional work is needed to fill in some missing pieces, several states have reviewed the Model Act and Agreement and enacted legislation to participate. In addition, multistate vendors are also reviewing the SSTP proposal to understand what it might mean for them should states in which they have customers (whether or not the vendor has nexus in the state) adopt the SSTP proposal.

Features of the SSTP proposal include state level administration of sales and use tax collections, uniformity in the state and local tax bases, a central electronic registration system, uniform sourcing rules, uniform definitions, uniform audit procedures, simplified tax returns, and consumer privacy protections.

Taxation Reports of Various Countries and the OECD

In addition to the United States, the OECD, Australia, Canada, the European Union, and a few other industrialized countries have issued extensive reports on e-commerce tax issues. Common themes in these reports include

- The Internet and e-commerce present opportunities for both governments and businesses.
- New taxes should not be imposed because restricting development of the Internet and e-commerce will only harm the country's economy.
- Neutrality should be considered in applying tax laws to transactions in e-commerce so that the tax law does not distort behavior.
- Multiple taxation must be avoided.
- Tax systems should be simple in order not to hinder a business's expansion of its market into the large markets offered through e-commerce.
- Countries will need to work together to deal with tax issues so as to avoid multiple taxation and undue competition, update and coordinate treaty provisions, coordinate the legal basis for taxing multinational transactions, and coordinate enforcement powers.
- Governments and taxpayers should work together to identify and address issues.

LOOKING AHEAD

E-commerce is a new business model for which existing tax systems were not designed. While some e-commerce transactions fit clearly within existing rules, many do not. The issues are complicated and global in nature. Efforts have been underway since at least 1996 to address these issues at the international level, as well as state and local levels. Work on identifying the specific issues and possible solutions will likely continue for many years due to the scope and range of issues. At the international level, the OECD is the likely avenue for resolving tax issues due to the purpose of the OECD and the depth of study it has performed to date. At the state and local level, states will continue to work on simplifying their sales and use tax system in the hopes of being able to collect sales tax from remote vendors. Unfortunately, the states do not have a strong record of working together to resolve issues and not all states are participating in the SSTP. As the state tax debate continues, so will the discussion of the proper role of Congress in helping to resolve the issues. Finally, although there has been little attention to them, a few

e-commerce issues at the federal level will likely be addressed in the next few years.

The issues and discussions present an opportunity for taxpayers, tax practitioners, and others to provide input to the debate to shape the future redesign or design of tax systems.

GLOSSARY

Consumption tax A tax on what a person consumes, generally based on the price charged for such items, in contrast to other types of taxes, such as an income tax, which is a tax on what a person earns. As with the federal income tax, which does not tax all types of income, a consumption tax may exclude certain types of items consumed. For example, many state sales tax systems exempt food. A consumption tax could be in the form of a *sales tax*, a consumed income tax (where the tax base is income less savings), or a value-added tax.

Nexus A connection between a taxpayer and a jurisdiction that allows the jurisdiction to subject the taxpayer to taxation without it being viewed as unfair. For state taxation, the degree of the connection required is dictated by the U.S. Constitution (Due Process and Commerce Clauses) and federal legislation.

OECD (Organization for Economic Cooperation and Development) An organization of industrialized countries that works to promote policies to achieve high economic growth and employment, rising living standards, financial stability, and expansion of world trade; has held major conferences on taxation and e-commerce and formed technical advisory groups (TAGs) of experts from government and businesses to discuss key e-commerce taxation issues.

Sales tax A tax imposed on the sale of property and services, generally imposed on retailers (and in some states, service providers) who are allowed to collect it from buyers (that is, to pass the tax onto buyers). In the United States, the District of Columbia and all states except for Alaska, Delaware, Montana, New Hampshire, and Oregon impose a sales tax. Most states impose the sales tax on tangible personal property and some services. The states have a variety of exemptions—property and services not subject to sales tax, such as certain clothing and food. The exemptions vary from state to state.

Use tax A tax complementing a sales tax and imposed by all U.S. jurisdictions that also impose a sales tax; generally applies when a taxpayer buys a taxable item outside of the state for use inside the state. For example, when a resident buys a taxable item (such as a book) from a remote vendor (one without a physical presence in the state), the resident is responsible for submitting the use tax to the state tax agency.

Value-added tax (VAT) A tax on the value added to goods and services at each stage of production and distribution. For example, if a manufacturer adds value to a raw good in the form of wages, the wages are subject to VAT. There are three methods that a jurisdiction may use to measure value added: credit invoice, subtraction, and addition. The most common method is the credit invoice that measures value added as sales less purchases (with the difference being the value added by the taxpayer who had the sales). The amount of tax collected under each method is the same and generally is the same amount as would be collected under a sales tax. A commonly cited benefit is that the tax is collected throughout the production and distribution process rather than only at the final retail sale stage (as is the case with a sales tax).

CROSS REFERENCES

See *Electronic Commerce and Electronic Business; Legal, Social and Ethical Issues; Public Accounting Firms.*

REFERENCES

Advisory Commission on Electronic Commerce (n.d.). Retrieved November 1, 2002, http://www.ecommercecommission.org/

American Institute of CPAs (AICPA) (2001). *Tax policy concept statement no. 1—Guiding principles of good tax policy: A framework for evaluating tax proposals.* Retrieved November 1, 2002, from http://ftp.aicpa.org/public/download/members/div/tax/3-01.pdf.

Cline, R. J., & Neubig, T. S. (1999). *The sky is not falling: Why state and local revenues were not significantly impacted by the Internet in 1998.* Washington, DC: Ernst & Young.

Council on State Taxation (COST) (2002). 2001 State study and report on telecommunications taxes. Special Report *9*(2). Washington, DC: Bureau of National Affairs.

European Union (1997). *A European Initiative in Electronic Commerce,* COM (97) 157. Retrieved November 1, 2002, from http://www.cordis.lu/esprit/src/ecomcom.htm

Federation of Tax Administrators (FTA) (1997). *Sales taxation of services: 1996 update* (Research Report No. 147). Washington, DC: Federation of Tax Administrators.

Government Accounting Office (GAO) (2000). *Sales taxes—Electronic commerce growth presents challenges: Revenue losses are uncertain* (GAO/GGD/OCE-00-165). Washington, DC: U.S. Government Accounting Office.

Mazerov, M., & Lav, I. J. (1998). *A federal "moratorium" on Internet commerce taxes would erode state and local revenues and shift burdens to lower-income households.* Retrieved November 1, 2002, from the Center on Budget and Policy Priorities Web site: http://www.cbpp.org/512webtax.htm

OECD (1998). *Electronic commerce: taxation framework conditions.* Retrieved November 1, 2002, from http://www.oecd.org/pdf/M000015000/M00015517.pdf

OECD (2000). *Clarification on the application of the permanent establishment definition in e-commerce: Changes to the commentary on the model tax convention on Article 5.* Retrieved November 1, 2002, from http://www.oecd.org/pdf/M000015000/M00015535.pdf.

OECD (2001). *Tax treaty characterization issues arising from e-commerce.* Retrieved November 1, 2002, from http://www.oecd.org/pdf/M000015000/M00015536.pdf.

OECD (n.d.) About: Tax and electronic commerce. Retrieved November 1, 2002, from http://www.oecd.org/EN/about/0,,EN-about-101-nodirectorate-no-no-no-29,00.html

Quill Corporation v. North Dakota, 504 U.S. 298 (1992).

Streamlined Sales Tax Project (SSTP) (n.d.). *Streamlined Sales Tax System for the 21st century.* Retrieved November 1, 2002, from http://www.geocities.com/streamlined2000/

U.S. Treasury (1996, November 21). *Selected tax policy implications of global electronic commerce.* Retrieved November 1, 2002, from http://www.ustreas.gov/offices/tax-policy/library/internet.html

U.S. Treasury (2002). *Statement by Deputy Treasury Secretary Kenneth W. Dam on European Union e-commerce tax proposal.* Retrieved November 1, 2002, from http://www.ustreas.gov/press/releases/po1001.htm

Washington State, Department of Revenue (1998). *Retailers' Cost of Collecting and Remitting Sales Tax,* Retrieved November 1, 2002, from http://dor.wa.gov/docs/reports/Retailers_Cost_Study/retailsum.htm

FURTHER READING

Hardesty, D. (2002). *E-commerce tax news.* Retrieved November 1, 2002, from http://www.ecommercetax.com/

Nellen, A. (2002). *E-commerce taxation links.* Retrieved November 1, 2002, from http://www.cob.sjsu.edu/facstaff/nellen_a/e-links.html

TCP/IP Suite

Prabhaker Mateti, *Wright State University*

INTRODUCTION

The Internet and the World Wide Web are based on TCP/IP. The term "TCP/IP" refers to not only the TCP (transmission control protocol) and IP (Internet protocol), but also includes other protocols, applications, and even the network medium. These protocols include UDP, ARP, and ICMP. These applications include telnet, FTP, Secure Shell, NFS, Web browsers and servers, and the many items collectively called the Web services. This chapter is an encyclopedic survey of these topics starting from the seven-layer OSI model to recent improvements in the implementations of the protocol stack and firewalls.

A computer system *communicates* with another system by sending a stream of bytes. A *byte* is a sequence of 8 bits. A *checksum* is the arithmetic sum of a sequence of numbers used to detect errors that may have altered some of the numbers in the sequence. The communication is actually between a process running on one system with one running on the other system. The two processes communicate information in a pre-agreed form known as protocol. That is, the two processes agree on the meaning of specific byte values occurring in specific positions in the stream.

This chapter describes the core protocols known as IP and TCP, and a few application protocols based on these. The details of IP and TCP are not directly experienced by the ordinary user unless a network sniffer is used. Nevertheless, it is crucial to understand these before attempting to understand the application protocols.

In each protocol, there is a stream of bytes known as a frame, a datagram, a packet or a segment depending on the "level." We describe the content of such a protocol data unit as a rectangular diagram such as the one shown in Fig. 2. The width of such a diagram is always 32-bits, numbered from 00 to 31. The units digits of these bit indices are shown in one row, and the tens digits are shown in the row above it. Each boxed row stands for a sequence of 32 bits (4 bytes).

LAYERS

Computer networking is easier to understand as a stack of layers, each layer providing the functionality needed by the layer above it. There are two such models.

The *OSI* (*Open Systems Interconnection*) *model* of computer networks has seven layers. Each layer provides functionality that the next higher layer depends on.

The *physical* layer provides the physical means of carrying the stream of bits. Ethernet, Fast Ethernet, Wireless 802.11, T-carrier, DSL (digital subscriber line), and ATM are examples of this layer. All media are considered functionally equivalent. The differences are in speed, convenience, and cost. Converters from one media to another exist and make it possible to have different physical layers in a computer network.

The *data link* layer takes the raw stream of bits of the physical layer and provides the functionality of sending and receiving a meaningful message unit called a *frame* and also provides error detection functions. A frame includes checksum, source and destination addresses, and data. The frame boundaries are special patterns of bits. Software of this layer will retransmit a frame if it is damaged, say due to a burst of noise on the physical layer. The data link layer is divided into the media access control

(MAC) sub-layer, which controls how a computer on the network gains access to the data and permission to transmit it, and the logical link control (LLC) sublayer, which controls frame synchronization, flow control, and error checking. This layer describes the specification of interface cards to specific types of networks, e.g., Ethernet, Token Ring, etc. Protocols from the TCP/IP suite that occupy this layer are SLIP and PPP.

The *network* layer accepts messages from the source host, converts them into packets of bytes, and sends them through the data link. This layer deals with how a route from the source to the destination is determined. This layer also deals with congestion control. The IP, address resolution protocol (ARP), reverse ARP (RARP), Internet control message protocol (ICMP), and IGMP belong to this layer.

The *transport* layer transfers data, and is responsible for end-to-end error recovery and flow control. The TCP and user datagram protocol (UDP) belong to this layer.

The *session* layer establishes, manages, and terminates connections between the programs on the two hosts that are communicating. The concepts of ports and connections belong to this layer.

The *presentation* layer provides independence from possibly different data representations of the host machines. The HTTP (hypertext transfer protocol), FTP (file transfer protocol), telnet, DNS, SNMP (simple network management protocol), NFS (network file system), etc. belong to this layer.

The *application* layer supports the end-user invoked programs. FTP, HTTP, IMAP (Internet message access protocol), NTP (network time protocol), POP3 (post office protocol version 3), rlogin (remote login), SMTP (simple mail transfer protocol), SNMP, SOCKS, telnet, X-Window, Web services, etc. are part of this layer.

The OSI model is officially recognized by the ISO (International Standards Organization). The practical world of TCP/IP networking was in full use by the time the OSI model was formulated. Its unofficial model, referred to as the *TCP/IP model,* the U.S. Department of Defense *DoD model,* or even more simply the *Internet model,* organizes the networks into the following five layers:

The *physical* layer, matching the OSI physical layer.

The *link* layer, similar to the OSI link layer. This is also called the network access layer. It defines the network hardware and device drivers.

The *IP* layer, also called the Internet protocol layer. This layer deals with identifying individual hosts and routing. It is similar in functionality to the OSI network layer.

The *transport* layer, containing the UDP and TCP. The UDP provides "connectionless service" and TCP provides "connection-oriented service." UDP does not guarantee, unlike the OSI transport layer, reliable delivery.

The *application* layer, handling the responsibilities of the session, presentation, and application layers of the OSI model.

The data unit of each layer is *encapsulated* by the layer below it. An application data unit [AP] is encapsulated by the TCP layer, which prefixes a TCP header TCPH as [TCPH [AP]]. The IP layer encapsulates it as [IPH [TCPH [AP]]], where IPH is the IP header. Assuming the Ethernet, it encapsulates it as [EH [IPH [TCPH [AP]]] FSC], where EH is the Ethernet header and FSC is a frame check sequence, generated by the Ethernet hardware.

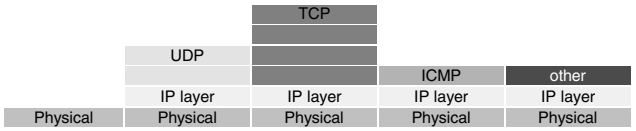


Figure 1: Protocol stack.

Protocol Stack

The computer network literature talks of protocol stacks. Each item in the stack is a layer of software that implements a collection of protocols. In Fig. 1, the relative heights indicate the level of functionality and the dependency of the layers. Except for the layer marked "Physical" all others are software layers. The protocol stack is an integral component in a modern operating system.

Lower Layers

In this section, we describe the layers that support the IP. The TCP/IP protocol family runs over a variety of network media including IEEE 802.3 (Ethernet) and 802.5 (Token Ring) LANs, X.25 lines, satellite links, and serial lines.

Ethernet

Ethernet can support IP and other protocols simultaneously. Ethernet was invented in 1972 at Xerox PARC by Metcalfe and his colleagues. They named it so as to emphasize the capability of the physical medium to carry bits to all hosts, analogous to the "luminiferous ether" of old physics. The word Ethernet now refers to 10 megabits-per-second (Mbps) transmission speed, Fast Ethernet refers to 100 Mbps, and Gigabit Ethernet refers to 1000 Mbps. The media varieties include the current twisted-pair (10baseT, with RJ45 connectors), the original thick coaxial system (10base5), thin coaxial (10base2), and fiber optic systems (10basesF). Ethernet has been standardized by the Institute of Electrical and Electronics Engineers as IEEE 802.3.

All hosts attached to an Ethernet are connected to a shared signaling medium. Ethernet signals are transmitted serially, one bit at a time, over the shared medium that every attached host can observe. Ethernet is a broadcast medium. That is, when a frame is sent out on the Ethernet, every controller on the local unswitched network can see the frame. To send data, a host waits for the channel to become idle, and transmits its frame. All hosts on the network contend equally for the transmission opportunity. Access to the shared medium is governed by the MAC mechanism based on the Carrier Sense Multiple Access with Collision Detection (CSMA/CD) system. This ensures that access to the network channel is fair, and that no single host can lock out other hosts. If two or more devices try to transmit at the same instant, a transmit collision is

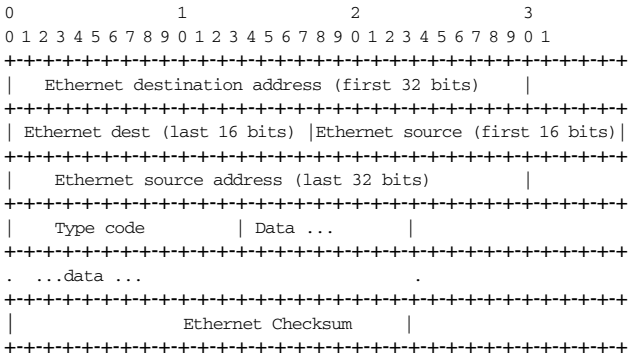


Figure 2: Ethernet frame.

detected, and the devices wait a random (but short) period before trying to transmit again.

An Ethernet controller comes with a 48-bit MAC address built-in from the factory. Every Ethernet frame has a 14-byte header that includes the Ethernet MAC addresses of the source and destination, and a 2-byte type code (see Fig. 2). The type code identifies the protocol family (such as IP, ARP, NetBEUI). The Data field is from 48 to 1500 bytes in length. Following the data, there is a checksum computed by the Ethernet controller for the entire frame.

The Ethernet address is written with each byte in hexadecimal, separating bytes with either a hyphen or a colon. Every device is expected to listen for Ethernet frames with their destination address. All devices also listen for Ethernet frames with a wild-card destination address of "FF-FF-FF-FF-FF-FF" (in hexadecimal), called a "broadcast" address. When these packets are received by the Ethernet Network Interface Card (NIC), it computes the checksum, and throws the packet away if an error is detected by the checksum. If the type code is IP, the Ethernet device driver passes the data portion of the frame up to the IP layer of the OS. Under OS control, the NIC can be put into a so-called *promiscuous* state wherein the NIC listens to all frames regardless of their destinations.

Asynchronous Transfer Mode (ATM)

Asynchronous transfer mode (ATM) is widely deployed as a backbone technology. ATM uses 53-byte fixed-length packets called cells for transport. Information is divided among these cells, transmitted, and then re-assembled at their final destination. ATM is connection-oriented. ATM itself consists of a series of layers. Its physical layer is based on various transmission media that range in speed from kilobits per second to gigabits per second. The layer, known as the adaptation layer, holds the bulk of the transmission. This 48-byte payload divides the data into different types. The ATM layer contains 5 bytes of additional information, referred to as overhead.

Serial Line Internet Protocol (SLIP)

A serial network is a link between two computers over a serial line, which can be a dial-up connection over telephone lines or a direct connection between the serial ports of two computers. Serial line IP (SLIP) [Request for Comment (RFC) 1055, available at http://www.rfc-editor.org/] defines the encapsulation protocol, just as an Ethernet frame envelopes an IP packet. Unlike Ethernet, SLIP supports only the IP, and not multiple protocols across a single link. The serial link is manually connected and configured, including the specification of the IP address. SLIP does not provide mechanisms for address negotiation, error correction, or compression. However, many SLIP implementations record the states of TCP connections at each end of the link, and use header compression that reduces the size of the combined IP and TCP headers from 40 to 8 bytes.

Point-to-Point Protocol (PPP)

Point-to-point protocol (PPP) [RFC 1661, RFC 2153] replaces the older SLIP, and is an encapsulating protocol for IP and other protocol datagrams over serial links. The encapsulation and framing adds 2, 4, or 8 bytes depending on the options chosen. PPP includes a link control protocol (LCP) that negotiates the encapsulation format, sizes of packets, authentication methods, and other configuration options. The CCP (compression control protocol) used by PPP negotiates encryption. The *IP control protocol* (*IPCP*) included in the PPP configures the IP address, and enables the IP protocol on both ends of the point-to-point link.

Point-to-Point Tunneling Protocol (PPTP)

Point-to-point tunneling protocol (PPTP) encapsulates PPP packets into IP datagrams. Its use is in providing virtual private networks (VPN). After the initial PPP connection to a PPTP server, a PPTP tunnel and a PPTP control connection are created. *Tunneling* is the process of sending packets of a certain protocol embedded in the packets of another protocol. PPTP uses an enhanced Generic Routing Encapsulation (GRE) mechanism to provide a flow- and congestion-controlled encapsulated datagram service for carrying PPP packets.

THE INTERNET PROTOCOL

IP [RFC 791] delivers a sequence of bytes from a source host S to a destination host D, even when the hosts are on different networks, geographically vastly separated. The byte sequence and the destination are given to the IP layer by the upper layer. The IP layer forms an IP datagram from the sequence. An IP datagram consists of an IP header followed by transport layer data. The IP layer software discovers routes that the packet must take from S to various intermediate nodes, known as routers, ultimately arriving at D. Thus, IP is routable. Each packet travels independently even when the host S wishes to send several packets to D; each packet delivery is made independently of the previous ones. The routes that each packet takes may change. Thus, IP is connectionless. The IP layer is designed deliberately not to concern itself with guaranteed delivery (i.e., packets may be lost or duplicated), but instead it is a "best effort" system. The ICMP described later aids in this effort.

IP Address

An operating system during boot-up assigns a unique 32-bit numeric ID known as its IP address to each NIC located in the host system. There is no rigid relationship between

the Ethernet address and the IP address. The IP address is obtained either by looking it up in a configuration file or via dynamic host configuration protocol (DHCP). The IP addresses are carefully controlled world-wide. The IANA, Internet Assigned Numbers Authority (www.iana.org), assigns the so-called public IP addresses to organizations and individuals upon application.

Three address ranges known as Class A, Class B, and Class C are of importance. In a Class A address, the 0-th bit is always a 0, bits 1 through 7 identify the network, and bits 8 through 31 identify the host, permitting 2^{24} hosts on the network. In a Class B address, the bit 0 is always a 1, bit 1 is always a 0, bits 2 through 15 identify the network, and bits 16 through 31 identify the host, permitting 2^{16} hosts on the network. In a Class C address, bits 0 and 1 are both 1 always, bit 2 is a 0 always, bits 3 through 23 identify the network, and bits 24 through 31 identify the host, permitting 2^8 hosts on the network.

IP addresses are typically written in a dotted-decimal notation as a.b.c.d where a is the first byte, b the second, c the third, and d the fourth byte. Each of a–d is a number in the range 0–255, written in the decimal notation. A subnet is a collection of hosts whose IP addresses match in several bits indicated by the ones in a sequence of 32-bits known as a subnet mask, also written in the dotted-decimal notation. Thus, 255.255.255.0 is a mask of 24 ones followed by 8 zeroes. Because of this structure, the mask is also written as /24. Nodes and routers use the mask to identify the address of the network on which the specific host resides. The address of the network is the bit-wise AND of the IP address and the mask. The host ID is the bit-wise AND of the IP address and the complement of the mask.

Occasionally, a network node X needs to discover certain information from other nodes, but the node X does not know the addresses of these others. In such situations, X broadcasts using special destination IP addresses. The direct broadcast address of X is the address whose host ID is all ones, and whose network address equals that of X. The limited broadcast address is 255.255.255.255.

The following three blocks of the IP address space is intended for private internets:

10.0.0.0 to 10.255.255.255 (10/8 prefix, Class A)

172.16.0.0 to 172.31.255.255 (172.16/12 prefix)

192.168.0.0 to 192.168.255.255 (192.168/16 prefix, Class C)

That is, on the Internet at large, there must never be IP packets whose source or destination addresses are from the above ranges.

Most operating systems are internally structured to depend on the presence of a network layer. To facilitate this, the address 127.0.0.1 is assigned as the so-called address of the localhost (spelled as one word) and 127.0.0.0 as the localnetwork (spelled as one word). Packets sent to this address do not actually travel onto the external network. They simply appear as received on the local (artificial) device.

When a machine is physically moved from one network to another, we must reassign an IP address that belongs to the new network. This is one of the problems that mobile IP solves.

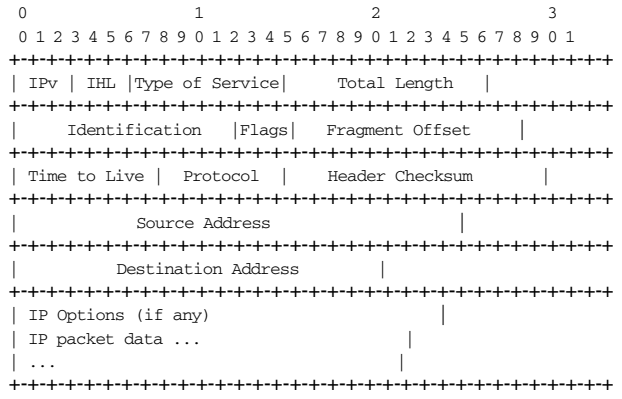


Figure 3: IP header.

IP Header

An IP header is a sequence of bytes that the IP layer software prefixes to the data it receives from the higher layers. The resulting IP header plus the data is given to the lower layer (e.g., the Ethernet card device driver). The byte layout of IP headers is shown in Fig. 3. The header may or may not have the IP Options filed. Except for this field, all other fields are fixed in length as shown. Minimally (i.e., without the options), the IP header is 20 bytes in length. With IP options, an IP header can be as long as 60 bytes.

IPv is the version number of the protocol; currently it is 4. IP version 6 is discussed later. The value of IHL multiplied by 4 is the length of the IP header in bytes. The Type of Service field specifies the "relative urgency" or importance of the packet. Total Length is a 2-byte field giving the length, in bytes, of the entire packet including the header, options (if any), and the packet data. Thus, the maximum length of an IP data gram is 65535 bytes. (IP over Ethernet limits this to 1500.) The Identification field, Flags, and Fragment Offset are used to keep track of the pieces when a datagram must be split up as it travels from one router to the next. IP fragmentation is discussed further below. The Time to live (TTL) is a number that is decremented by 1 whenever the datagram passes through a router node. When it goes to 0, the datagram is discarded, and an error message is sent back to the source of this packet. The Protocol field identifies the protocol of the data area. The Header Checksum field is the 16-bit one's complement of the one's complement arithmetic sum of the entire header viewed as a sequence of 16-bit integers. The Source Address is the datagram's sender IP address, and Destination Address is the IP address of the intended recipient. The IP Options may or may not be present. When present, its size can be one or more bytes.

Routing Protocols

When the source S and the destination D are on the same network, we have a direct delivery of the packet that does not involve routers. When the two hosts are not on the same network, in general, there can be multiple paths between the source and destination. Because of failures, maintenance, and other reasons, the intermediate nodes

known as routers may come on or off during the delivery of packets. Thus, consecutive packets sent by a host S to a destination D may (have to) travel entirely disjoint routes depending on how the network is connected. The typical network host has only one NIC, and hence is on only one network, and sending and receiving of network traffic is secondary to its main functionality. Routers are specialized computer systems whose primary function (often their sole function) is to route network traffic. Routers must have multiple NICs, each on a separate network. A router examines the destination IP address of a packet and consults its routing tables that record information regarding where to deliver a packet next so that definite progress is made in moving the packet closer to its final destination.

Every network host (including routers) has a routing table, which can be visualized as a table of two columns: To send the packet to a final destination given in column 1, send the packet to the next hop whose IP address is given in column 2. The size of such a table can be greatly reduced by parametrizing the column 1 by its network address, and also by including a default row in the table that acts as a catch-all. The default row indicates the next hop IP address for any packet whose destination network address does no match that of any other row. Once the next hop IP address is determined, the router uses the lower layer address (such as the Ethernet MAC) to deliver the packet to the next hop.

The routing table of an ordinary host rarely changes from boot-up to shut down. The tables of routers, however, must be dynamic and adjust to changing conditions, perhaps by the millisecond, of the Internet. Routing protocols keep the routing tables up-to-date.

The Internet is a network of autonomous networks. Interior gateway protocols (IGPs) maintain the routing tables within an autonomous network. RIP (routing information protocol) and OSPF (open shortest path first) are examples of IGPs. Border gateway protocol (BGP) is the most common protocol in use for routing among autonomous networks.

IP Fragments

When datagrams are too large to be sent in a single IP packet, due to interface hardware limitations for example, they can be split up by an intermediate router unless prohibited by the Don't Fragment flag. IP fragmentation occurs when a router receives a packet larger than the Maximum Transmission Unit (MTU) of the next network segment. All such fragments will have the same Identification field value, and the Fragment Offset indicates the position of the current fragment in the context of the pre-split-up packet. Intermediate routers are not expected to re-assemble the fragments. The final destination will reassemble all the fragments of an IP packet and pass it to higher protocol layers (like TCP or UDP).

Domain Name Service

Because of the mnemonic value, humans prefer to work with host names such as gamma.cs.wright.edu. A host name in this form is known as a fully qualified domain name (FQDN); gamma is the name of the host, and cs.wright.edu is the name of the domain the host is in. Each network host maintains a short cache table of FQDNs to IP addresses. When a name is not found in this cache, the host enquires with a name server the domain name service (DNS) protocol. Each name server behaves recursively in this manner. Sometimes it is necessary to transfer the resource records of an entire DNS zone. A DNS query with Name = wright.edu, Class = IN, and Type = AXFR will trigger a zone transfer for all the host names in the wright.edu domain.

DNS uses a distributed database to delegate control of domain name hierarchies among zones, each managed by a group of name servers. Name servers are the repositories of information that make up the domain database. Each name server has authoritative information about one or more zones, but may also have cached, but non-authoritative, data about other parts of the database. The name server marks its responses to queries as authoritative or not.

Either TCP or UDP can be used for DNS, connecting to server port 53. Ordinary DNS requests can be made with TCP, although convention dictates the use of UDP for normal operation.

Mobile IP

As the mobile network host moves, its point of attachment may change, and yet in order to maintain existing transport-layer connections, it must keep its IP address the same.

The *mobile node* uses two IP addresses. The *home address* is static and is used to identify TCP connections. The *care-of address* changes at each new point of attachment. Whenever the mobile node moves, it *registers* its new care-of address with its home agent. The home agent redirects the packets to the current care-of address by constructing a new IP header that contains the care-of address as the destination IP address. This new header encapsulates the original packet, causing the home address to have no effect on the routing of the encapsulated packet until it arrives at the care-of address. When the packet arrives at the care-of address, the effect of this "tunneling" is reversed so that the packet once again appears to have the home address as the destination IP address.

Mobile IP discovery of the care-of address uses an existing standard protocol called Router Advertisement (RFC 1256). A router advertisement carries information about default routers, and in addition carries further information about one or more care-of addresses. Home agents and care-of agents typically broadcast these advertisements at regular intervals (say, once every few seconds). If a mobile node needs to get a care-of address in a hurry, it multicasts a router solicitation. An advertisement also informs the mobile node whether the agent is a home agent, a care-of agent, or both, and therefore whether it is on its home network or a care-of network, and about special features provided by care-of agents, for example, alternative encapsulation techniques.

The registration of the new care-of address begins when the mobile node, possibly with the assistance of the care-of agent, sends a registration request to the home address. The home agent typically updates its routing table.

Registration requests contain parameters and flags that characterize the tunnel through which the home agent will deliver packets to the care-of address. The triplet of the home address, care-of address, and registration lifetime is called a *binding* for the mobile node. The home agent authenticates that registration was originated by the mobile node.

Each mobile node and home agent compute an unforgeable digital signature using one-way hash algorithm MD5 (Message Digest 5, RFC 1321) with 128-bit keys on the registration message, which includes either a time stamp or a random number carefully generated.

Occasionally a mobile node cannot contact its home agent. The mobile node tries to register with another home agent by using a directed broadcast IP address instead of the home agent's IP address as the target for the registration request.

TRANSMISSION CONTROL PROTOCOL

TCP [RFC 793, RFC 3168] offers the client process a connection to a server process. This connection needs to be established, as needed. Once this connection is established, the TCP protocol guarantees the correct (both in content and in order) delivery of the data. TCP sends its message content over the IP layer, and can detect and recover from errors. TCP, however, does not guarantee any speed of delivery, even though it offers congestion control.

Ports and Connections

Port numbers are used by the transport layer for multiplex communication between several pairs of processes. To each message, this layer adds addresses, called port-numbers. The port numbers would have been assigned by the OS to certain processes. Thus, a connection is uniquely identified by four numbers: source and destination IP addresses, and source and destination port numbers. The IP addresses are supplied by the IP layer. The TCP and UDP port numbers are unrelated to the memory addresses often referred to as IO ports.

Reliable Transmission

TCP requires that every segment include an acknowledgment of the last data segment received in the other direction. TCP is a sliding window protocol with time-out and retransmits. If the sender does not receive an acknowledgment within the time-out period, it retransmits the segment. Acknowledgments are piggybacked on reply data. There is dynamically adjustable window size that specifies the number of bytes the receiver has as buffer space. The sender continues to send and slides the window ahead as long as acknowledgments are being received for bytes within the window.

TCP messages, called segments, are sent as one or more IP datagrams. A TCP header follows the IP header, supplying information specific to the TCP protocol. Figure 4 contains the details of the TCP segment.

The letters |U|A|P|R|S|F| in the fourth row of the segment are abbreviated names for control bit flags: URG, Urgent Pointer field significant, ACK, Acknowledgment field significant; PSH, Push Function; RST, Reset the

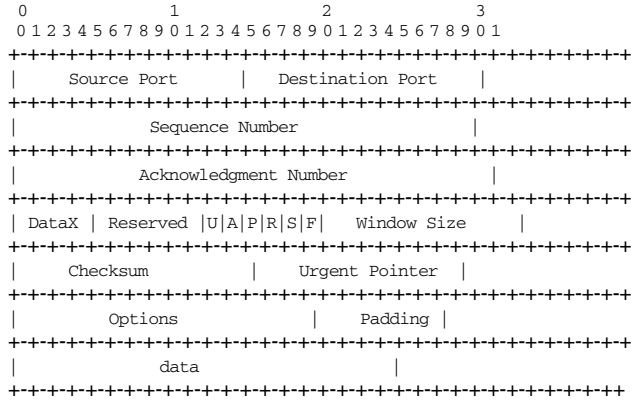


Figure 4: TCP header.

connection; SYN, Synchronize sequence numbers; and FIN, sender is finished with this connection. These are further explained below.

The Sequence Number, together with the Acknowledgment Number, serves as a ruler for the sliding window protocol. While establishing the connection, the SYN flag is set to 1, and the client and the server exchange their initial sequence numbers.

Acknowledgment Number is valid only when the ACK bit is set. This field contains the value of the next sequence number the sender of the segment is expecting to receive. Once a connection is established, this is always included. The DataX number multiplied by 4 is the number of bytes in the TCP Header. This indicates where the data begin. Window size is described in the section Congestion Control. Urgent Pointer is valid when URG is 1. Its value is a positive offset from the sequence number in this segment. Options, if any, are given at the end of the TCP header and are always a multiple of 8 bits in length. All options are included in the checksum. An option can be just a single byte, or it can be a byte of option-kind, followed by a byte of option-length, and the actual option-data bytes. The option-length counts the two bytes of option-kind and option-length as well as the option-data bytes.

State Diagram

A TCP server process starts its life by passively opening a port and starts to listen to connection attempts from clients. This process causes a number of changes in the information maintained by the TCP layer software. These transitions are described by the state diagram shown in Fig. 5. An active open causes a SYN = 1 segment to be sent out and the software enters the SYN-sent state. Below we describe two handshakes that establish a connection and close a connection.

TCP Three-Way Handshake

This establishes the connection between the initiating node (say A, the client) and the receiving node (say B, the server) of packets as follows:

A: "I would like to talk to you, B." A sends a packet with SYN = 1, and the initial sequence number to B.

B: "OK, let's talk." B replies with a SYN–ACK packet (i.e., SYN = 1, ACK = 1, Acknowledgment number =

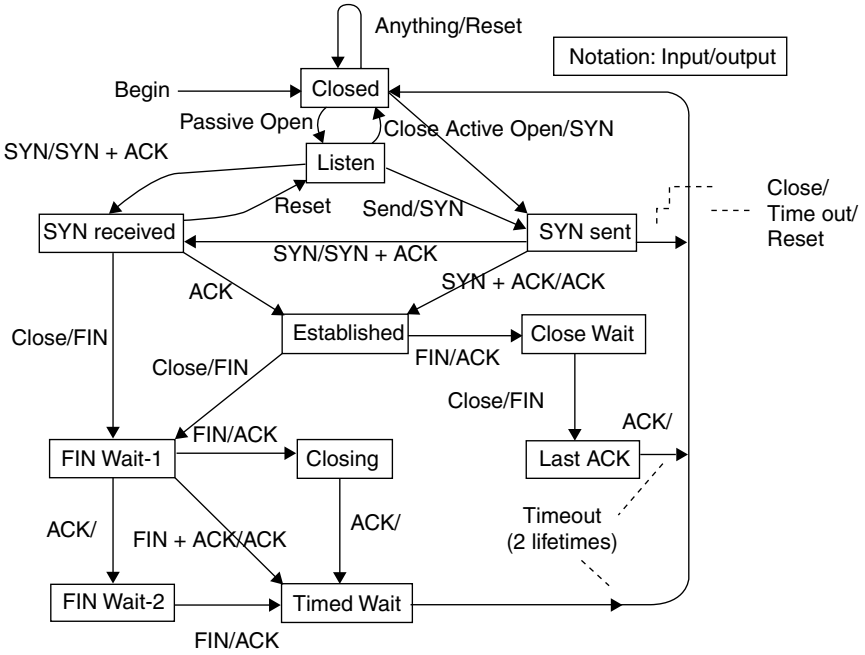


Figure 5: TCP state diagram.

sequence number received + 1, and SYN = 1 with sequence number set to the ISN of the server. If B was unwilling, it responds with a RST = 1 packet refusing the request for service.

A: "Thanks!" A sends a packet with ACK = 1, Acknowledgment number = ISN of B + 1, SYN = 0, sequence number = previous sequence number + 1.

The sequence number for SYN = 1 can be a zero, but that is not secure, so the sequence number is randomly chosen. Here is an example:

SYN	ACK	src	dst	Sequence-number	Acknowledgement-number
1	0	1037	80	102723769	0
1	1	80	1037	1527857206	102723770
0	1	1037	80	102723770	1527857207

where the client is on port 1037 establishing a connection with a service on port 80 (typically HTTP).

Four-Way Handshake

This terminates a previously established connection between say A and B as follows:

A sends to B a packet with FIN = 1, which indicates "no more data from A." This flag is used when closing a connection down the normal way. The receiving host B enters the CLOSEWAIT state and starts the process of gracefully closing the connection. Each end of the connection sends a packet with the FIN = 1. The receiver is expected to acknowledge a received FIN packet by sending a FIN = 1 packet.

B sends to A a packet with ACK = 1, acknowledging the FIN packet received.

B sends to A another packet, but now with FIN = 1.

A sends to B a packet with ACK = 1. No further packets are exchanged.

So four packets are used to close a TCP connection in the normal situation.

Closing a connection can also be done by using the RST flag set to 1, which indicates to the receiver that a reset should occur. The receiving host accepts the RST packet provided the sequence number is correct, and enters the CLOSED state and frees any resource associated with this instance of the connection. The RST packet is not acknowledged. A host H sends a connection resetting RST packet if host X requested a connection to a non-existent port p on host H, or for whatever reason (idle for a long time, or an abnormal condition, etc.), the host H (client or the sever) wishes to close the connection. Resetting is unilateral. Any new incoming packets for that connection will be dropped.

TCP Timers

TCP depends on many timers.

Connection Establishment Timer is started when the SYN is sent during the initial connection setup. Typical value of this timer is 75 s. If a time-out occurs, the connection is aborted.

FIN_WAIT timer is started when there is a transition from the *FIN_WAIT_1* state to the *FIN_WAIT_2* state. The initial value of this timer is 10 min. If a packet with FIN = 1 is received, the timer is canceled. On expiration of the 10 min., the timer is restarted with a value of 75 s. The connection is dropped if no FIN packet arrives within this period.

TIME_WAIT timer is started when the connection enters the *TIMED-WAIT* state. This is to allow all the segments in transit to be removed from the network. The value of the timer is usually set to 2 min. On expiration of the timer, the connection is terminated.

For the KEEP_ALIVE timer, we need to distinguish the silence caused because there are no data to send from

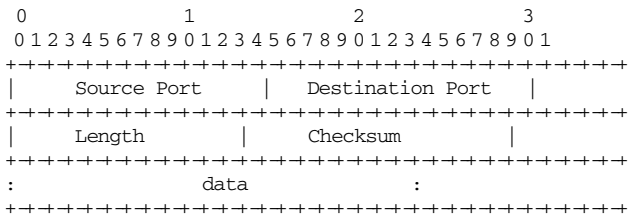


Figure 6: UDP header.

that caused by a broken connection. Setting a keep-alive timer allows TCP to periodically probe the other end. The default value of this timer is 2 h. After the expiration of the timer, probes are sent to the remote end. The connection is dropped if the remote does not respond.

Congestion Control

The Acknowledgment segments from a receiver also contain an advertised window size that specifies the buffer size it has available. If the new size is larger, the sender can increase; if it smaller, the sender can decrease the sliding window size. TCP uses the sliding window mechanism to control the flow.

When a router begins to accumulate too many packets, it can send ICMP Source Quench messages to the senders of these packets. These messages should cause the rate of packet transmission to be slowed.

UDP, ICMP, AND OTHER PROTOCOLS

User Datagram Protocol (UDP)

UDP is a connectionless transport protocol. It is a thin protocol on top of IP, providing high speed but low functionality. Delivery of UDP datagrams is not guaranteed. Nor can it detect duplicate datagrams. The UDP protocol is mostly used by application services where squeezing the best performance out of existing IP network is necessary such as trivial file transfer (TFTP) and NFS, and by the DNS.

The port numbers appearing in the UDP header are similar to the TCP port numbers (see Fig. 6), but the OS support required by UDP ports is much simpler and less resource consuming than that of TCP ports. Source Port is the port of the sending process. When not meaningful, this field is set to 0. Destination Port is the UDP port on the receiving machine, whose IP address is supplied by the IP layer. Length is the number of bytes in the datagram including the UDP header and the data. Checksum is the 16-bit one's complement of the one's complement sum of the UDP header, the source and destination IP addresses obtained from the IP header, and the data, padded with zero bytes at the end (if necessary) to make a multiple of 2 bytes.

Internet Control Message Protocol (ICMP)

ICMP [RFC 792, 1981] manages and controls the IP layer, as in reporting network errors, such as a host or entire portion of the network being unreachable or a packet being directed at a closed port, reporting network congestion, assisting in trouble shooting, reporting time-outs, or forcing routing options. In general, much of the best-effort in delivering IP datagrams is associated with ICMP. The purpose of the ICMP messages is to provide feedback and suggestions about problems, for example, when a datagram cannot reach its destination, when the gateway does not have the buffering capacity to forward a datagram, or when the gateway can direct the host to send traffic on a shorter route. To avoid the infinite regress, no ICMP messages are sent about ICMP messages. Also ICMP messages are only sent about errors in handling fragment zero of fragmented datagrams.

An ICMP message is sent as the data portion of an IP datagram. These IP datagrams are treated like all other IP datagrams. Each ICMP message begins with a 1-byte ICMP type field, which determines the format of the remaining data, a one-byte code field, and a 2-byte checksum. If the ICMP message is reporting an error, these 4 bytes are followed by the first 8 bytes of the IP datagram causing the error.

The popular network utilities ping and traceroute use ICMP. The ping command sends several echo requests, captures their echo replies, and displays statistics about speed and datagram loss. The traceroute utility constructs IP datagrams with well-chosen TTL values and collects the time-exceeded ICMP messages to map a route from the source to a destination IP address.

ICMP helps improve the performance of the network. For example, ICMP redirect messages from a router inform a host that a different router is more optimal than it is for certain destinations.

Address Resolution Protocol (ARP)

ARP [RFC 826, 1982] is used to determine the Ethernet MAC address of a device whose IP address is known. This needs to be done only for outgoing IP packets, because IP datagram must be Ethernet framed with the destination hardware address. The translation is performed with a table look-up.

Reverse ARP (RARP) [RFC 903] allows a host wishing to discover its own IP address to broadcast its Ethernet address, and expect a server to reply with its IP address.

The ARP cache accumulates as the host continues to network (see Table 1). If the ARP cache does not have an entry for an IP address, the outgoing IP packet is queued, and an ARP request packet that effectively requests "If your IP address matches this target IP address, then please let me know what your Ethernet address is" is broadcast. Once the table is updated as a result of

Table 1 A Small Portion of an ARP Cache

IP address	Ethernet address
130.108.2.23	08-00-69-05-28-99
130.108.2.1	00-10-2f-fe-c4-00
130.108.2.27	08-00-69–0d-99-12
130.108.2.20	08-00-69-11-cf-b9
130.108.2.10	00-60-cf-21–2c-4b
192.168.17.221	00-50-ba-5f-85-56
192.168.17.112	00-a0-c5-e5–7c-6e

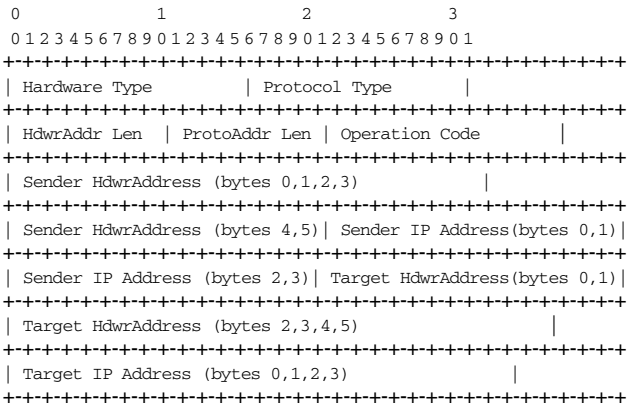


Figure 7: An ARP request/response packet.

receiving a response, all the queued IP packets can now be sent.

The entries in the table expire after a set time period in order to account for possible hardware address changes for the same IP address. This change may have happened, e.g., due to the NIC being replaced.

ARP is an OSI layer-3 protocol, but it does not use an IP header. It has its own packet format as shown in Fig. 7. The ARP request packet has zeroes in the target Hardware Address fields. It is broadcast on the local LAN without needing to be routed. The destination host sends back an ARP reply with its hardware address so that the IP datagram can now be forwarded to it by the router. An ARP response packet has the Sender/Target field contents swapped as compared to the request.

TCP/IP SECURITY

The TCP/IP suite had many design weaknesses so far as security and privacy are concerned, all perhaps due to the era (1980s) that the development took place. For example, the ICMP redirect message, intended to improve routing performance, has often been used maliciously. In this section, we summarize some of these issues. All major OS have made improvements in their implementations of the protocol stack that disable many of the attacks described below.

Covert Channels

A number of protocols permit covert channels. For example, ICMP echo request packets should have an 8-byte header and a 56-byte payload. ICMP echo requests should not be carrying any data. However, such ICMP packets can be significantly larger, carrying covert data in their payloads. Covert channels are prevalent in nearly all the protocols of the TCP/IP suite.

IP Address Spoofing

IP spoofing replaces the IP address of (usually) the sender or (in rare cases) the destination with a different address. Because the IP layer of the OS normally adds these IP addresses to a data packet, a spoofer must circumvent the IP layer and talk directly to the raw network device. IP spoofing is normally used to deposit another exploit on the target machine.

Note that the attacker cannot simply reassign the IP address of T to the host A using ifconfig or a similar configuration tool. Other hosts, as well as T, will discover (through ARP, for example) that there are two machines with the same IP address.

IP Fragment Attacks

Many firewalls do not perform packet reassembly. Attackers create artificially fragmented packets in order to fool such firewalls.

A well-behaving set of IP fragments is non-overlapping. However, a cleverly constructed second fragment packet can have an offset value that is less than the length of the data in the first fragment, so that upon packet reassembly it overrides several bytes of the first fragment.

In the IP layer implementations of nearly all OS, there are bugs in the reassembly code. An attacker can create and send a pair of carefully crafted but malformed IP packets that in the process of reassembly cause a server to "panic" and crash.

The Ping of Death attack sends fragments that when reassembled will be a packet larger than the maximum permissible length.

TCP Flags

TCP segments have a number of flags that have, collectively, a strong influence on how the segment is processed. However, not all the flags can be independently set or reset. For example, SYN FIN, SYN FIN PSH, SYN FIN RST, SYN FIN RST PSH are all illegal combinations. Past implementations have accounted only for valid combinations, ignoring the invalid combinations as simply "will not happen." Attackers have written special programs that construct such illegal packets, and cause the network hosts to crash or hang.

The SYN Flood

In the TCP protocol as designed, there is no limit set on the time to wait after receiving the SYN. An attacker initiates many connection requests with spoofed source addresses to the victim machine. The SYN+ACK packets that the victim host sends are not replied to. Once the limit of the half-open connections is reached, the victim host will refuse further connection establishment attempts from any host until a partially opened connection in the queue is completed or times out. This effectively removes a host from the network for several seconds, making it useful at least as a stepping tool to other attacks, like IP spoofing.

TCP Sequence Number Prediction

TCP exploits are typically based on IP spoofing and sequence number prediction. In establishing a TCP connection, both the server and the client generate an initial sequence number from which they will start counting the packets transmitted. This sequence number is (should be) generated at random, and should be hard to predict. However, some implementations of the TCP/IP protocol

make it rather easy to predict this sequence number. The attacker either sniffs the current SEQ/ACK of the connection or can algorithmically predict them.

Closing a Connection by FIN

The attacker constructs a spoofed FIN packet. It will have the correct SEQ numbers so that it is accepted by the targeted host connection. This host would believe the (spoofed) sender did not have any data left. Any packets that may follow would be ignored as bogus. The rest of the four-way handshake is also supplied by the attacker.

A similar connection killing attack using RST has also been seen.

Connection Hijacking

Host YY accepts the packets from XX only when correct SEQ/ACK numbers are used. The attacker ZZ can send one or two packets to YY spoofing the source address as XX, at a time when XX was silent. YY would accept these data, and update ACK numbers. XX would continue to send its old SEQ numbers, as it is unaware of the spoofed data. As a result, subsequent packets from XX are discarded by YY. The attacker ZZ could then impersonate to be XX, but using correct SEQ/ACK numbers from the perspective of YY. This results in ZZ hijacking the connection: host XX is confused while YY thinks nothing is wrong as ZZ sends "correctly synchronized" packets to YY.

APPLICATIONS

Nearly all network applications are based on a client–server architecture where one process, the client, requests services from a second process, the server. Typically, the client and server processes are on different machines, but they need not be.

File Transfer Protocol (FTP), Telnet, and rlogin

The three application protocols described in this section are all based on TCP. They send authentication information and data in the clear (i.e., unencrypted), and hence are easily compromised by network sniffers. Also, their authentication of host is simply the IP address that they respond to. Consequently, utilities based on these protocols should not be used in situations where security is a concern. The SSH described later provides near equivalent functionality at a higher level of security.

Telnet

Telnet [RFC 854] establishes a TCP connection with a telnet server on the reserved port 23, and passes the keystrokes of the telnet client to the server, and accepts the output of the server as characters to be displayed on the client. The server presents these keystrokes as input received from a pseudo-terminal to the OS hosting the telnet server. Telnet defines a network virtual terminal (NVT) format as that which permits interoperability with machines that use different characters for common operations such as terminating a line and interrupting a run-away process. The telnet client typically maps the signal-generating keys of the keyboard to invoke the corresponding control functions of the NVT. The control functions are encoded as escape sequences of 2 bytes, the IAC (255), followed by the 1-byte code of the control function. Telnet uses the URGENT DATA mechanism of TCP to send control functions so that the telnet server can respond appropriately.

File Transfer Protocol (FTP)

FTP [RFC 959, 1985] uses two TCP connections, one called the *control* connection and the other the *data* connection. The client can issue a number of commands on the control connection that change various settings of the FTP session. All content transfer occurs on the data connection. The FTP client opens a control connection to port 21 of the FTP server machine. This connection persists the entire session. The format of data passed over the control connection is the same as that of telnet NVT. The GET command requests for the transfer of the contents that the server has (popularly known as *downloading*), and the PUT command requests the server to receive and store the contents that the client is about to send (popularly known as *uploading*).

The data connection can be opened in two modes. In the *active mode* FTP, the server initiates a data connection as needed from its port 20 to a port whose number is supplied by the client via the PORT command. In the *passive mode* FTP, the server informs the client a port number higher than 1024, to which the client initiates a data connection.

rlogin

The rlogin protocol [RFC 1282] is similar in functionality to telnet, and also operates by opening a TCP connection on the rlogin server machine at port 513. It is widely used between UNIX hosts because it provides transport of more of the UNIX terminal environment semantics than does the telnet protocol, and because on many UNIX hosts it can be configured not to require user entry of passwords when connections originate from trusted hosts.

Dynamic Host Configuration Protocol (DHCP)

DHCP [RFC 2131, 1997] consists of a protocol for delivering host-specific configuration parameters from a DHCP server to a host, and a mechanism for allocation of IP addresses to hosts. The IP configuration parameters that DHCP can supply include subnet mask, a list of default routers, TTL, and MTU. A typical host will use DHCP soon after booting into the OS to configure its network. DHCP assumes that the IP layer software will pass the packets delivered to the NIC of the host even though the IP address has not been assigned yet. DHCP has three mechanisms for IP address allocation. In "automatic allocation," DHCP assigns a permanent IP address to a client. In "dynamic allocation," DHCP leases an IP address to a client for a limited period of time (or until the client explicitly relinquishes the address). In "manual allocation," a client's IP address is manually assigned but uses DHCP to convey the assigned address to the client. Dynamic allocation is the only one of the three mechanisms that allows automatic reuse of an address that is no longer needed by the client to which it was assigned.

Hypertext Transfer Protocol (HTTP)

HTTP [RFC 2616, 1999] is at the core of the World Wide Web. The Web browser on a user's machine and the Web server on a machine somewhere on the Internet communicate via HTTP using TCP usually at port 80. HTTPS [RFC 2660, 1999] is a secure version of HTTP.

A Web browser displays a file of marked-up text with embedded commands following the syntactic requirements of the hypertext markup language (HTML). There are several ways of invoking these commands, the most common one being the mouse click. Most of the clickable links displayed by a Web browser are the so-called links that associate a URL (universal resource locators) with a visible piece of text or graphic. URLs have the following syntax: scheme://[userName[:password@]] serverMachineName[:port]/[path][/resource][?parm1 = parma&parm2 = parmb]. A simple example of the above is http://www.cs.wright.edu/~pmateti/InternetSecurity where the scheme was chosen to be http, the port defaults to 80, and the path given is ~pmateti/InternetSecurity. A click on such a link generates a request message from the browser process to the Web server process running on the remote machine whose name www.cs.wright.edu is obtained from the link clicked.

HTTP Message Format

The request and response are created according to the HTTP message format, which happens to be a sequence of lines of text. The first line identifies the message as a request or response. The subsequent lines are known as header lines until an empty line is reached. Following the empty line are lines that constitute the "entity body." The header lines have a left-hand side that names various parameters separated from the right-hand side that provides values with a colon. The request line has three components: a method (one of GET, POST, or HEAD), a URL, and the version number of HTTP (either 1.0 or 1.1) that the client understands. Of the methods GET is the most common. The POST method is used when the client sends data obtained from a user-filled HTML form. The HEAD method is used in program development. The response line also contains three components: HTTP/version-number, a status code (such as the infamous 404), and a phrase (such as Not Found, OK, or Bad Request). The entity body in a response message is the data, such as the content of a Web page or an image, that the server sends.

Authentication and Cookies

Web servers requiring user authentication send a WWW-Authenticate: header. The Web client prompts the user for a username and password, and sends this information in each of the subsequent request messages to the server. HTTP is stateless in that the HTTP server does not act differently to request based on previous requests. Occasionally, a Web service wishes to maintain a minor amount of historical record of previous requests. Cookies [RFC 2965] create a stateful session with HTTP requests and responses. The response from a server can contain a header line such as "Set-cookie: value." The client then creates a cookie stored on the client's storage. In subsequent requests sent to the same server, the client includes the header line "Cookie: value."

Secure Shell (SSH)

SSH provides the functionality of telnet and rlogin but with greater security. The user name and password are sent encrypted after establishing a TCP connection on port 22, authenticating that the connection is indeed to the server. The SSH client maintains a database of server names and their authentication keys that the server offers the first time an SSH session is opened to the server. All subsequent SSH sessions compare the authentication key offered by the server with that stored in the client database. The SSH provides for other methods of authentication.

CONCLUSION

The Internet and the World Wide Web are based on a suite of protocols collectively known as TCP/IP. It includes not only the transmission control protocol and Internet protocol, but also other protocols such as UDP, ARP, and ICMP, and applications such as telnet, FTP, Secure Shell, and Web browsers and servers. We surveyed these topics starting from the seven-layer OSI model to recent improvements in the implementations of the protocol stack and firewalls.

GLOSSARY

Big endian The lowest address of a 4-byte-long integer that could be occupied by the most significant byte as agreed upon by the two machines that are communicating.

Byte A sequence of 8 bits used by one computer system to communicate with another by sending several in a stream. Viewed as a number, it is in the range of 0 to 255.

Checksum The arithmetic sum of a sequence of numbers used to detect errors that may have altered some of the numbers in the sequence.

Client A process that establishes connections for the purpose of sending requests.

Connections In the connectionless communication one process sends data to another without prior negotiation. The recipient does not acknowledge the receipt of the message, and the sender has no guarantee that the message is indeed delivered. In the connection-oriented communication there are three well-defined phases: connection establishment, data transfer, and connection release.

Datagram A sequence of bytes that constitutes the unit of transmission in the network layer (such as IP).

Direct link Connects two hosts often by multiple paths as these paths may change over time as short as a few milliseconds.

Frame The unit of transmission at the data link layer, which may include a header and/or a trailer, along with some number of units of data.

Host A device capable of sending and receiving data over a network; often, a computer system with a network interface card (NIC), but it can be a much simpler device.

Little endian The highest address of a 4-byte-long integer that could be occupied by the most significant byte as agreed upon by the two machines that are communicating.

Network A collection of links in which the hosts are connected either directly or indirectly.

Network applications Programs that operate over a network.

Network operating systems Operating systems with network software built-in and are that aware of byte order issues.

Node Synonym for host.

Octet A byte on older computer architectures where the smallest addressable unit of memory was a word and not a byte.

Packet A generic term used to designate any unit of data passed between communicating entities, and is usually mapped to a frame.

Process The dynamic entity that can be summarized as a "program during its execution on a computer system."

Program A file of binary data in a certain rigid format that is specific to each platform, capable of being both a client and a server (our use of these terms refers only to the role being performed by the program for a particular connection, rather than to the program's capabilities in general).

Protocol A formal and pre-agreed set of rules that govern the communications between two or more entities.

Request for Comments (RFC) documents The collection of Internet standards, proposed designs, and solutions published by researchers from universities and corporations soliciting feedback and archived at http://www.rfc-editor.org/

Server A process that accepts connections in order to service requests by sending back responses, also known as a daemon.

Tunneling The process of sending packets of a certain protocol embedded in the packets of another protocol.

CROSS REFERENCES

See *Circuit, Message, and Packet Switching; Internet Security Standards; Standards and Protocols in Data Communications; Virtual Private Networks: Internet Protocol (IP) Based.*

FURTHER READING

TCP/IP details are part of many college courses on computer networks. There are several textbooks. Of these, the three authoritative volumes of Comer's *Internetworking with TCP/IP* are classic technical references in the field aimed at the computer professional and the degree student. Volume I surveys TCP/IP, and covers details of ARP, RARP, IP, TCP, UDP, RIP, DHCP, OSPF, and others. There are errata at http://www.cs.purdue.edu/homes/dec/tcpip1.errata.html. *The Internet Book: Everything You Need to Know about Computer Net working and How the Internet Works* is a gentler intro duction. The books listed above by Tanenbaum, and Kurose and Ross are also popular textbooks. The book by Stevens discusses from a programming point-of-view.

Routing protocols are discussed briefly in the above books. The books by Halabi, and Kurose and Ross cover this topic very extensively.

The HTTP protocol and related issues are thoroughly discussed in the books of Krishnamurthy and Rexford, and Gourley and Totty.

The book by Denning and Denning is a high-level discussion of how the vulnerabilities in computer networks are affecting society. Mateti has an extensive Web site (http://www.cs.wright.edu/~pmateti/InternetSecurity) that has lab experiments and readings online. The book by Garfinkel and Spafford explores security from a practical UNIX systems view.

The Usenet newsgroup comp.protocols.tcp-ip is an active group and maintains a frequently asked questions (FAQ) document that is worthwhile reading. The Technical Committee on Computer Communications of the IEEE Web site (http://www.comsoc.org/) maintains an extensive collection of conference listings. The *IEEE/ACM Transactions on Networking* is a peer-reviewed archival journal that publishes research articles.

Comer, D. (2000a). *Internetworking with TCP/IP Volume 1: Principles, protocols, and architecture* (4th ed.). Englewood Cliffs, NJ: Prentice Hall.

Comer, D. (2000b). *The Internet book: Everything you need to know about computer networking and how the Internet works* (3rd ed.). Englewood Cliffs, NJ: Prentice Hall.

Denning, D. E., & Denning, P. J. (1998). *Internet besieged: Countering cyberspace scofflaws*. Reading, MA: Addison Wesley.

Garfinkel, S., Spafford, G., & Schwartz, A. (2003). *Practical Unix and Internet Security* (3rd ed.). Sebastapol, CA: O'Reilly.

Gourley, D., & Totty, B. (2002). *HTTP: The definitive guide*. Sebastapol, CA: O'Reilly.

Halabi, B. (2000). *Internet routing architectures*. Indianapolis, IN: Cisco Press.

Hunt, C. (2002). *TCP/IP network administration* (3rd ed.). Sebastopol, CA: O'Reilly.

Iren, S., Amer, P. D., & Conrad, P. T. (1999). The transport layer: tutorial and survey. *ACM Computing Surveys, 31,* 360–404.

Krishnamurthy, B., & Rexford, J. (2001). *Web Protocols and Practice: HTTP/1.1, networking protocols, caching, and traffic measurement*. Reading, MA: Addison Wesley.

Kurose, J. F., & Ross, K. W. (2003). *Computer networking: A top-down approach featuring the Internet* (2nd ed.). Reading, MA: Addison Wesley.

Mateti, P. (2002). Internet security class notes. Retrieved (Oct 2002) from www.cs.wright.edu/~pmateti/InternetSecurity.

Stallings, W. (2003). *Cryptography and network security: Principles and practice* (3rd ed.). Englewood Cliffs, NJ: Prentice Hall.

Stevens, W. R. (1996). *TCP/IP Illustrated, Vol. 3: TCP for transactions, HTTP, NNTP, and the UNIX domain protocols*. Reading, MA: Addison Wesley.

Stewart, J. W., III. (1999). *BGP4: Inter-domain routing in the Internet*. Reading, MA: Addison Wesley.

Tanenbaum, A. S. (2003). *Computer networks* (4th ed.). Englewood Cliffs, NJ: Prentice Hall.

RFCs are archived at http://www.rfc-editor.org/

Telecommuting and Telework

Ralph D. Westfall, *California State Polytechnic University, Pomona*

INTRODUCTION

Telecommuting is a fascinating concept, with a history that dates back to fiction published nearly 100 years ago. As a concept, it has achieved a high level of publicity, and most people know (or think they know) what it involves. Actual usage of telecommuting to avoid driving to work is still relatively low, however, and may never attain the levels forecasted by pundits such as Toffler (1980) in his discussion of "the electronic cottage."

This chapter first looks at the history of the telecommuting concept. The subsequent section discusses definitions of telecommuting and various other kinds of telework. These definitions are critically important to understanding telework, because the various types have different implications for individuals, organizations, and the larger society. The definitions are followed by a discussion of the implications of the varieties of telework. Following this is an analysis and discussion of the two key measures of telecommuting activity: the participation or proportion of the working population that telecommutes and the rate or frequency of telecommuting as a substitute for driving to the office. This is followed by a critical analysis of reported productivity gains from telecommuting.

The subsequent section is a discussion of usage trends and forecasts, which is followed by a discussion of the impact of the Internet and other technological trends on telecommuting. A discussion of the implications of theories from economics and sociology on telecommuting usage and recommendations for organizations and employees interested in telecommuting concludes the chapter.

A BRIEF HISTORY OF TELECOMMUTING AND TELEWORK

Telecommuting is a fascinating concept. People think about it when driving through congested traffic on their way to work. It sounds appealing to persons who are chafing at the restrictions of being in a specified place for a set number of hours almost every working day, often with much less control than they would like over what they are doing and how they are doing it.

The appeal of this concept is reinforced by its association with other concepts that have positive implications. Telecommuting is touted as a new, different, and better way of working. It is associated with computing and networking technologies that seemed advanced when the concept started receiving popular attention, and these technologies have become ubiquitous in recent years. It promises other benefits, including reductions in air pollution and dependence of imported petroleum supplies, by reducing the number of cars on the freeways. Telecommuting sounds like a solution to many of the ills of modern society—noise, congestion, air pollution, crowded living conditions—that have been noted since the industrial revolution.

Although telecommuting had been discussed before in publications with limited circulation, the concept received its first widespread publicity in 1980 in *The Third Wave* by Toffler. This book devoted a whole chapter to telework and became a best seller in the United States and around the world. Toffler suggested that 10 to 20% of the workforce could be working from their homes (described as "electronic cottages") within 20 to 30 years and that this work at home would be comparable to the prevalent work mode before the industrial revolution led to the relocation of labor to factory settings (Toffler, 1980).

Many of the concepts of telecommuting had been around long before Toffler. In view of the appealing aspects of the concept and the long-standing concerns about urbanization and congestion, it should not be a great surprise that many of these ideas, including the virtual office, were the background for a science fiction piece published back in 1909, *The Machine Stops*. The author of this short story was E. M. Forster, the British writer whose novels include *Howard's End* and *A Room with a View*.

The setting of *The Machine Stops* (Forster, 1909) is a future where people interact with each other almost exclusively through networking technologies. They use a telephone, at the same time viewing a live image of the other person in a glowing round plate that they can hold in their hands. They deliver or attend lectures using these same media. When they become ill, doctors attend to them through what we now describe as telemedicine. In

contrast to the predominantly favorable current publicity on the topic of virtual work and virtual social relationships, however, Forster's vision is of an anti-utopian way of life, isolated from other people and the external world.

Although not central to the story as in *The Machine Stops,* communications technologies that are similar to ones now used for telecommuting are also mentioned in H. G. Wells' science fiction novel, *When the Sleeper Wakes* (Wells, 1899). In this novel, a man falls asleep around 1900, awakens 200 years later, and becomes a messianic figure in this new environment in which more advanced technologies are available.

Norbert Wiener, who coined the word "cybernetics," touched briefly on the technologies of telecommuting in a book published in the middle of the 20th century (Wiener, 1950). He suggested that an architect in Europe could supervise the construction of a building in the United States, using the "Ultrafax" to send drawings to the United States and photographs of the building under construction back to Europe. The architect could also use the telephone and teletype for other communications regarding the project.

In the following decade, Frederick Memmott, a New York state transportation analyst, published a serious conceptual discussion of the benefits of telecommuting. He noted that the transfer of information was the cause of much urban transport activity and therefore could be handled instead by information and communications technologies of that period (telephone, closed-circuit television, and radio). He stated that the primary function of a business executive is communication and expressed his view that executives therefore could use these technologies to handle all of their responsibilities from their homes or other locations relatively close to their homes (Memmott, 1963).

The late 1960s and 1970s spawned additional research and discussion of telecommuting and related activities. Harkness (1973) completed a Ph.D. dissertation on the topic of communications-enabled decentralization of office space in urban centers. Several universities implemented distance-learning programs using closed-circuit television and telephone links to students in off-site locations. Hospitals in the Boston area used high-resolution television equipment, including close-up lenses, for medical consultations or "telemedicine."

The Telecommunications-Transportation Tradeoff (Nilles, Carlson, Gray, & Hanneman, 1976) published the results of government-funded research into possible solutions to the "energy crisis" of the early 1970s. It included a detailed analyses of the feasibility, costs, benefits, human relations factors, and so forth of using telecommunications to relocate insurance processing employees from high-cost, downtown office space to suburban locations and also included analyses of distance learning, telemedicine, and other implementations. Nilles, the lead author of this book, actually coined the word *telecommuting*.

Since then, telecommuting has been the topic of hundreds of articles in both popular and academic periodicals and numerous books (a search on the Amazon.com Web site found more than 100 publications with the word *telecommuting* in their titles or keywords in 2003). Although these publications generally don't mention Toffler's predictions, which obviously are not being realized, they often cite more recent analyses that are almost as optimistic.

DEFINING TELECOMMUTING AND TELEWORK

Telecommuting is broadly applied to a range of somewhat different activities. It is important to identify these types of work because they have differing implications for policies and investments at both the governmental (regulatory and infrastructure) and organizational levels, as well as for other stakeholders including individual employees and the larger society. These definitions have a major impact on reported statistics and trends in the number of people who are identified as telecommuters. Very broad definitions can lead to high estimates that can be misleading if used without making distinctions among (a) the different types of activities and (b) differing usage levels (measured as a proportion of total working hours). Some researchers prefer to use the word *telework* (work at a distance), which actually encompasses most of the activities in a more precise fashion. Common usage, however, is to stay with the word telecommuting.

Telecommuting implies substituting telecommunications technologies for physical transportation. The slogan is "Bring the work to the worker, rather than bringing the worker to the work." An employee working all day at home, rather than *commuting* (usually alone in an automobile) to a work location, is the epitome of this concept. The remainder of this chapter identifies this type of telework as "classic telecommuting," whether done an average of once a week, several days a week, or every day ("full-time telecommuting").

Consistent with this popular view, the seminal research study on telework is the aforementioned *Telecommunications–Transportation Tradeoff* (Nilles et al., 1972). Much of the analysis in the book is related to decentralization of central city employees to suburban office locations, however, rather than having them work at home. Telecommunications technologies enable decentralized workers to access data in mainframe computers at the downtown offices, and local managers can also communicate with their off-site superiors largely by telephone. Although this kind of implementation does not eliminate commuting between home and the office, it can substantially reduce distances traveled. Similarly, the first "telecommuting dissertation" (Harkness, 1973) was actually an analysis of urban planning options to decentralize office work to nodes located outside the central business districts of major metropolitan areas.

"Telecommuting centers" represent another example of decentralization rather than complete substitution of telecommunications for computing. Local communities sponsored such centers, with the assistance of federal transportation funding, and private businesses have also offered them on a for-profit basis. There were approximately 30 such facilities in and around major urban areas in the state of California during the mid-1990s. They offered space on a subsidized basis, and some large organizations rented cubicles with the idea of having different

employees occupy the space on different days. Many of these centers were shut down soon after the initial subsidies phased down (or out), because of low usage of the facilities (Mokhtarian, 1996), and most of the rest are now gone.

"After-hours telecommuting" refers to employees using computers and telecommunications equipment, almost always in their own homes, to perform organizational work after (and possibly before) typical office hours. The word telecommuting is arguably a misnomer here, because this telework does not reduce traffic volumes on the transportation system.

Some employees may be able to use their home-computer capabilities to rationalize reducing or changing their hours in the office, thereby shifting their morning or evening commuting (or both) away from the most congested times. Such "part-day telecommuting" can reduce commuting times for individual employees, as well as reducing congestion and air pollution resulting from slower moving vehicles during the critical peak hours. Arrangements like this could also help organizations comply with traffic mitigation requirements. This is contingent on the requirements recognizing this approach, however, and on the incentives or sanctions being sizable enough to compensate organizations for the reduced face-to-face availability of some of their employees.

To the extent that they use information and communications technologies (ICT) to keep in contact with their organizations, people who work primarily at different locations or operate transportation vehicles are teleworkers. Such "mobile workers" include field sales and customer service representatives, transportation operatives (e.g., truck and cab drivers), consultants working at client facilities, and so on. There are other workers who have offices in traditional organizational locations but actually average 20% or more of their time outside of their buildings. Many of these employees also use information and communications technologies extensively when out of their offices and thus represent part-time mobile workers. The technologies mobile workers use are similar to those used by conventional telecommuters, except that wireless communications usually replace hardwired networks, and miniaturized devices (cell phones, personal digital assistants, laptop computers) are often used in place of their larger desktop counterparts.

Self-employed workers who work out of their homes really don't commute at all and thus should not be counted as telecommuters. Some of these "home workers" use the Internet extensively for research or communications with suppliers whose products or services they use to generate their own value-added outputs for their customers. Others are contract employees (e.g., computer programmers) who may work some assignments on site but handle others in a home worker mode by using the Internet to interface with the computers of the companies they work for. Therefore, home workers who use computers and the Internet in their activities are included among the ranks of teleworkers even though they don't really telecommute.

Electronic commerce represents another form of substitution of information and communications technologies for transportation. Although not a form of telework, e-commerce delivers goods to customers by mail or package delivery services. Thus e-commerce can replace vehicle trips—typically in low-occupancy vehicles going to one or a few scattered stops—with deliveries by carriers that service large numbers of locations via efficient routing. Customers handling banking transactions via home computers rather than traveling to their branches also reduces vehicle trips. Thus, increasing use of e-commerce also provides traffic and emission-reduction benefits.

The phrase *virtual organization* is applied to organizations that interface components of separate and distinct organizations for varying lengths of time to achieve specific objectives. The use of information and communications technologies makes it possible to coordinate between the organizations with less travel than would otherwise be required. In terms of surface vehicle travel, however, such reductions are typically small. For such ventures, the more important impact of information and communications technologies is to enable profitable alliances that otherwise would be more difficult because of the distances involved (e.g., short-term international joint ventures).

IMPLICATIONS OF TELECOMMUTING DEFINITIONS

Transportation planners and legislators are most interested in forms of telework that reduce peak-hour vehicle traffic volumes, either by reducing the amount of commuting ("classic telecommuting"), or by shifting some of it to off-peak times ("part-day telecommuting"). Reducing traffic obviously reduces air pollution and dependence on imported petroleum. Commuting at off-peak hours also helps somewhat, through the increased efficiency of automobile engines as a result of less stop-and-go driving.

On the other hand, organizations typically do not encourage either of these approaches because they reduce the proportion of regular working hours that office workers spend in their buildings. They are often willing, however, to allow one or both of these types of telecommuting on a case-by-case basis, where they make it possible to achieve other important objectives. A common rationale is retaining a valuable employee who might otherwise be unable to continue to work for the organization because of health issues or a relocation of either the family or the organization.

Organizations are the most interested in telework when it helps attain financial and other objectives. This typically occurs in situations in which, because of the requirements of their jobs, employees must work outside of an office to handle a significant portion or most of their responsibilities. In many "mobile work" situations, organizations have no choice but to allow off-site work and are therefore concerned with maintaining optimal communications with their remote employees, for example, for scheduling work, routing employees most efficiently to reduce costs, improving response times, and so forth.

For many information workers, after-hours telecommuting is becoming an unwritten requirement of their jobs. Unlike classic or part-day telecommuting, it does not reduce the face-to-face accessibility of employees during

regular working hours. Organizations also like after-hours telecommuting because it increases responsiveness to rapidly developing situations, generates additional hours of work from their employees, and provides supplemental disaster recovery capabilities. All these benefits come at negligible cost to many organizations, because much of the after-hours work is done by employees using hardware, software, and telecommunications services that they obtain at their own expense. (Some employees acquire these capabilities on favorable terms because of discounts negotiated by their organizations, e.g., acquiring computers or high-bandwidth telecommunications services.)

Classic telecommuting appeals to some employees. The benefits, especially reductions in the costs, stress, and time associated with driving on congested roads, are obvious. The decreased social contact, however, and perceptions that telecommuting will affect their prospects for advancement (or continued employment in downsizing organizations) are discouraging factors.

In the past, some advocates suggested "classic telecommuting" as a viable alternative to day care for working mothers. The magazine of the Conference Board had a cover picture in 1987 of a woman seated at a keyboard and answering a telephone while her baby peacefully napped in the background. The work of two researchers (Christensen, 1988; Olson, 1987) has shown that telecommuting was not a practical substitute for day care in typical situations. Classic telecommuting could be an option for employees with responsibilities for care of aging relatives or family members with disabilities, but the viability of such arrangements is obviously dependent on the amount and type of care required.

Classic telecommuting is sometimes mentioned as a possible means for integrating persons with disabilities into the workforce. There are problems with this concept. It may be unrealistic to suggest that people with little or no work experience can develop into productive workers away from the economies of scale and support available in an on-site setting. Some types or degrees of disability severely limit employment options, which could be a major factor for many who are disabled to the extent that commuting would not be practical. These issues might be less of a problem for persons who had significant work experience before their disabilities, but the potential loss of benefits when taking employment represents a discouraging factor. Although there have been pilot studies on telecommuting in relation to disabilities, reports on outcomes have not been widely distributed.

Among all the stakeholders, information and communications technology vendors are unquestionably the most enthusiastic advocates of telework. It increases demand for their hardware, software, and telecommunications services. The impacts are greatest for their more lucrative high-end offerings, such as laptop computers, wireless devices, high-bandwidth wired services (e.g., DSL or cable connections), and sophisticated communications software. Organizational consultants who advocate telecommuting also have strong economic interests in the topic, and these stakes could impair the objectivity of the research that some of them publish.

Table 1
Impacts of Varying Amounts and Rates of Telecommuting

PARTICIPATION (% OF EMPLOYEES)	USAGE RATE (% OF DAYS)	VOLUME REDUCTIONS (% TRAFFIC, REAL ESTATE USAGE)
10	20	2
20	10	2
20	5	1

TELECOMMUTING USAGE FACTORS

Two key factors predominantly determine the impact of classic telecommuting at both the societal and organizational levels. One is the amount of participation in telecommuting, the proportion of employees who regularly substitute telecommunications for travel. The other is the rate of telecommuting, how often employees make this substitution on the average. These two factors are multiplicative, as shown in Table 1.

The rates in Table 1 were not chosen arbitrarily: the first row represents 1 employee in 10 (10% participation) telecommuting 1 day/week (rate of 20%) on the average. This is typical of data from a number of studies of telecommuting implementations. The second and third lines represent more participation in telecommuting but at a lower rate. Some of the more optimistic estimates of participation count as telecommuters people who avoid a trip to the office as little as 1 day per month (3% rate). Usage rates vary between organizations, between organizational units, by type of work, and by geography. Therefore the rates (20% × 10%, and 20% × 5%) in the second and third lines in the table are actually representative of some situations.

The 20% usage rate seems to be relatively stable. Because classic telecommuting is done in increments of 1 day, this suggests that individuals and organizations are most comfortable with 1-day-per-week arrangements. Many knowledge workers handle a broad variety of tasks in typical weeks, and the mix varies at different times of the year. Some tasks are much more suited to telecommuting than others. Telecommuters can concentrate on suitable tasks at home for 1 day and leave the remainder of their tasks, ones that can be done equally well or better in the office, for other days in the week. The 1 day usually does not have to be on the same day every week, which provides additional flexibility relative to both the temporal distribution of tasks and the telecommuter's responsibilities to adjust to other employees' requirements. Being absent only 1 day means that telecommuters still have plenty of opportunities to interact with superiors, peers, and subordinates on a face-to-face basis and can minimize the inconveniences for others when they are not in the office.

The climate in some organizations, or in specific organizational units, may be unsupportive for telecommuting. In other situations, often dependent on the attitudes of individual managers, telecommuting is tolerated or even encouraged. Situations specific to certain organizations,

especially limited availability of real estate, may force an organization to encourage telecommuting as a temporary expedient. Therefore the participation rate fluctuates more than the usage rate.

Another limiting factor is that some employees frankly don't like to telecommute. In this regard, there is a "catch-22" consideration. Organizations typically reserve telecommuting for employees who have demonstrated good attendance and performance on the job. Having developed the habit of working hard 5 days of the week in an office, they may be less interested in—or actually uncomfortable with—being away from the office for even 1 day per week.

A researcher who has studied telecommuting extensively, and who is respected for the quality of her research, has developed a model of the participation and usage of telecommuting (Mokhtarian, 1998) based on analyses of her own and other published research data. The model starts with total employment, rather than just information workers as in other models, because some information workers are not able to telecommute and some employees who are not information workers are able to telecommute at least part of the time. This total is then adjusted for the percent who are able to telecommute (estimated at 16% at the date of the research, based largely on job suitability and manager willingness, and eventually rising to 30%). Of those who are able to telecommute, her analysis of research data indicates that only 50% will want to telecommute. Even after taking this into account some people—who could telecommute and would like to do so—will choose not to do so because of other concerns (leading to another adjustment of 76%). Multiplying all these factors together (16% × 50% × 76%) leads to an estimate of only 6% for the participation rate at the time of the study, and Mokhtarian projects that this will eventually rise to 11–12%.

As demonstrated in Table 1, the volume reductions generated by representative participation and usage rates are low, around 2% or less. This means that in most urban centers, telecommuting produces at best a hardly noticeable reduction in the aggregate load on the transportation infrastructure. There are concerns that telecommuters tend to move to locations or live in locations that are more remote, thus offsetting the telecommuting benefits by longer commutes on days that they do drive to their offices. Although studies on the issue have produced contradictory findings, the low participation rates mean that the impacts are negligible even if this is true.

The calculated volume reductions in Table 1 are not big enough to generate real estate savings for employees engaged in classic telecommuting. Some organizations have been able to achieve substantial real estate savings for teleworkers by reconfiguring the office facilities for employees in the mobile-worker category. Their organizations provide a limited number of cubicles or offices, which are used on a reservation basis ("hoteling"), or first-come, first-served ("free address"). These facilities are used by employees such as consultants, salespersons, and service workers who generally spend more of their working hours on the road or at client facilities than in offices in their own organizations. Thus the real estate savings do not result from a shift toward telecommuting; instead the savings result from a recognition that such employees do not require their own permanently assigned office space for 40+ hours per week.

TELECOMMUTING PRODUCTIVITY

Despite the enduring popularity of the concept, interest at the personal and societal level is not enough to lead to telecommuting at levels high enough to generate significant societal and environmental benefits. Telecommuting is contrary to the general trend of centralization of employees that started in the industrial revolution. In particular, it conflicts with the dominant employment pattern for information workers that has developed and evolved since the advent of high-rise office buildings more than 100 years ago.

Except for the self-employed, whether a person telecommutes or not is usually determined by their immediate supervisor or organizational policies (or both). Therefore, to achieve usage levels that are sufficient to produce significant environmental benefits, telecommuting needs strong support at the organizational level.

Researchers and writers on this topic point to a potential benefit that could provide a compelling rationale for the necessary organizational support. They claim that telecommuting (or telework) results in dramatic increases productivity.

A literature search on the impacts of telecommuting on productivity found claims of productivity gains of 30%, 35%, 43%, 50%, 65%, and 144% for groups of employees, and one statement that "many [employees] recorded 200% increases in output." Although some of these items were in articles by freelance writers in trade publications, the 35% and 50% figures came from a *Business Week* article ("It's Rush Hour," 1984). The claims have been made so often, from so many different sources, that a group of European researchers (Huws, Korte, & Robinson, 1990) noted a "surprising degree of unanimity" on this issue.

The claims of large increases in productivity are not limited to anecdotal reports. Nilles has made similar claims. His discussion of the findings of his research on a major telecommuting implementation in a governmental organization in the early 1990s included the somewhat surprising statement: "*37% of the work being accomplished in 18% to 23% of work week; possibly an average 100% productivity increase per telecommuting day*" (Nilles, 1994, italics in original).

In his economic analyses of the project benefits, Nilles used a supposedly conservative 22% productivity gain based on the average of managers' subjective impressions of employee productivity gains. Because the employees were telecommuting approximately 1 day per week, however, this implies a productivity gain of around 100% on the telecommuting day to achieve a 20% productivity gain for the week. Table 2 demonstrates this visually. (An alternate explanation is that telecommuting also increases productivity on nontelecommuting days, but a rationale for such gains is not readily apparent.)

As demonstrated in Table 2, even if telecommuting does dramatically improve productivity, the gains would be restricted in many cases by the typically low proportion of days per week that telecommuters work at

Table 2 Basis for a 20 Percent per Week Productivity Gain from One Day per Week Telecommuting

	TELECOMMUTING	IN OFFICE	TOTAL FOR WEEK
Actual hours	8	32	40
Productivity equivalent (hours)	16	32	48
Productivity gain/period (%)	100	0	20

home. A survey of pilot programs (Mokhtarian, 1996) indicated an average telecommuting rate of 1.2 days per week in 1991, which is comparable to the rates in the Nilles study. Another survey of more than 80 published studies (Bailey & Kurland, 2002) found similarly low rates.

Deconstructing Telecommuting Productivity

One way to evaluate the likelihood of potential productivity gains of large magnitude is to apply logic and common sense. Figure 1 shows a model of productivity based on four major factors:

1. Amount of work—actual hours of work, per day, week, month or year
2. Intensity of work—how hard the person is working
3. Efficiency of work—ratio of outputs to labor inputs (affected by amount of supporting technology, experience and training, organization of work, concentration, etc.)
4. Adjustments—telecommuters generally require additional inputs from the organization compared with other employees, and any such costs need to be netted out of productivity calculations

Based on this model a 10% increase in any factor, while the others remain constant or balance out, results in a productivity gain of a bit less than 10% (depending on the magnitude of the adjustments). A simultaneous 10% increase in two factors could produce a gain approaching 21%, and so on. The question then becomes this: What types of changes in these factors are likely to occur as a result of telecommuting?

Hours or Amount of Work

The average one-way commute in major urban areas is between 20 and 30 minutes per day. Thus, for an employee who works an eight-hour day, commuting represents around 10% of the work time. Assuming that the average telecommuter puts all the commuting time savings into extra work, that the level of intensity and efficiency do not decline, and that the adjustments are small, this would result in an increase in productivity of close to 10%. There are no guarantees that employees will actually devote all the extra time to more work than they would do in the office. From an employee relations point of view, telecommuting consultants warn that it would be bad policy to suggest or even imply that telecommuters should work extra hours.

Productivity = Hours x Intensity x Efficiency - Adjustments

Figure 1: Productivity model.

Another frequently mentioned source of extra work hours is recapturing time that would otherwise be lost because of medical appointments or problems in the home. For example, instead of losing a whole day because of a medical appointment near home, the employee could put in couple of hours at home rather than wasting the rest of the workday. (This argument assumes that all appointments are around the middle of the day, rather than in the early morning or late afternoon.) However simple calculations indicate a productivity gain of only 2.5% for a person who has a relatively high average of one such situation per month (12 situations × 4 hours/1900 work hours per year).

Intensity of Work

The telecommuting literature consistently emphasizes this aspect. Authors claim that, away from the distractions of the office, people will be able concentrate much better and get much more work done. There is anecdotal evidence for this in certain situations. For example Kidder's (1981) book, *The Soul of a New Machine,* mentioned a software engineer who ducked out to the Boston Public Library and developed the microcode for 195 minicomputer machine instructions in a very short time.

It is certainly true that people can work very intensely for short periods when highly motivated. The issue here, however, is the impact of telecommuting. Will this change in working circumstances motivate or energize people to work more intensely, on a sustained basis over months and years?

Telecommuting advocates suggest several possible mechanisms for increases in intensity. Not having to commute, telecommuters may have more energy to put into their tasks. If this extra energy is used to work longer hours, as discussed earlier, it may not be available to also increase the intensity of work.

Another argument is that some employees work better at certain times of the day, and these times may not coincide with traditional office hours. But even if there is a substantial proportion of the population that works significantly better at times that are largely outside traditional working hours, potential gains may be limited by the need for some to interact with other employees by telephone during the 8-to-5 time frame.

Efficiency of Work

People who make more use of information technology in their work are more productive, and by its very nature telecommuting requires more use of IT. Telecommuters in formalized programs usually receive extra training in using technologies and managing their work. In many formal programs, managers of telecommuters also receive additional training. Employees that are selected or allowed to telecommute typically have more experience and a track record of performance. Therefore, telecommuters should be more productive. These gains are not necessarily a result of telecommuting, however, and might be obtainable by providing comparable technology and training for other workers who remain on-site.

Another efficiency issue relates to the telecommuter's distance from the workplace. The modern office is an institution that has been evolving for more than 100 years. It provides efficient access to support personnel, high-speed office equipment, office supplies, and also to the paper files that are still present in many organizations. Telecommuters need to put in more time in planning to make sure all the necessary resources are available when away from work. They may need to devote extra time or money to obtaining some of these resources if they become necessary while telecommuting. Therefore for some aspects of their work, telecommuters may be less efficient than their counterparts back in the office.

Adjustments

Productivity gains need to be reduced by tangible expenses for equipment, technology support, training, telecommunications and other services, for example, that are greater than that received by nontelecommuters. In addition, the productivity calculations need to include intangible costs such as the following:

- Extra managerial supervision
- Extra support from other employees (e.g., faxing materials)
- Work transferred to other employees when telecommuters are unavailable
- Decreases in the effectiveness of meetings due to scheduling problems or less effective communication by remote participants

Hourly rates can be applied to these items to generate cost estimates. These intangibles may also create problems with morale and communications within an organization, however (Prusak & Cohen, 2001). The impacts of such adverse organizational impacts could exceed the actual labor costs to support telecommuters.

The business press abounds with anecdotal accounts of about increased productivity resulting from telecommuting. Published research studies are relatively rare, and many of these have not been of high quality. Like the study by Nilles (1994) these studies typically rely on subjective estimates of productivity. Westfall (1997) identified 15 possible factors that have not been adequately taken into account in much of the research on telecommuting productivity, including the subjective factors of "Hawthorne" effects and employees' inability to objectively evaluate their own work.

In this context, it should be noted that questions about the adequacy of the supporting evidence for productivity gains are not a secret among telecommuting researchers. Such concerns have been around for a number of years. For example, Kraut (1989, p. 20) stated that "It is not yet possible in most studies, however, to untangle the effects of novelty, self selection and longer work hours from the effects of work location." A more current review (Bailey & Kurland, 2002) reviewed more than 80 published studies and found that "little clear evidence exists that telework increases job satisfaction and productivity, as it is often asserted to do."

To put the productivity issue in context, consider that the continuing emphasis on increasing productivity throughout the United States and world economies has been a major driving force for investments in information technology. If telecommuting were able to generate noticeable productivity gains that flowed down to the bottom line, it would be reasonable to expect that companies with high proportions of knowledge workers would massively adopt telecommuting and continue to use it on a large scale. The fact that this has not happened is a telling indicator that telecommuting does not deliver, at least at the level of the whole organization, the productivity gains touted by consultants.

As an example of the hype on this topic, also consider the following anecdote. The author attended a seminar on telecommuting just after a major earthquake disrupted the freeway system in Southern California in 1994. One vendor mentioned that, in addition to providing relief from the horrendous traffic congestion at the time, telecommuting also offered "tremendous productivity gains." When asked why the details of these gains were not being published so that these benefits could become more widely known, the vendor said that organizations were using it as a "secret weapon."

TELECOMMUTING USAGE TRENDS

Toffler (1980) predicted that 10 to 20% of the population might be working in an "electronic cottage" mode by the year 2000. Although he did not define how much time people would have to work at home to fall into this category, his comparisons to the work environment before the industrial revolution suggests that more than half of the working hours for such workers would be at home. It is obvious that, based on any reasonable interpretation of the term, this prophecy has not come to pass (incidentally Toffler included little or no material on this subject in subsequent books).

Toffler was by no means the only person who offered optimistic predictions of growth in telecommuting. A cursory search of literature published during the 1990s finds numerous citations of forecasts of high rates of growth in telecommuting, and subsequent surveys predicting continued growth from higher usage levels. These forecasts were developed by commercial firms, however, not published in sources that were publicly available at reasonable costs and not subject to academic review of the data or methodologies. Analyses of the limited data published

on methodologies shows that the high levels reported in later surveys are based on extremely broad definitions of telecommuting (e.g., as low as 1 day per month), and a failure to distinguish between people who actually substitute telecommunications for physical travel versus other teleworkers (e.g., mobile workers, home-based self-employed workers, etc.). (The author's e-mailed requests for additional details on two such forecasts did not receive responses.)

THE INTERNET AND OTHER TECHNOLOGICAL TRENDS FAVORING TELEWORK

In 1969, Gordon Moore, one of the founders of Intel, noted that the number of transistors per chip required for the most efficient production of integrated circuits seemed to be doubling every year. He later revised his estimate to 18 months, and this observation came to be known as "Moore's Law." The concept is popularly understood as meaning that the power of computers doubles every 18 months (or sometimes 2 years), whereas the cost remains constant.

The progress of the computer industry seems to be tracking Moore's Law well, resulting in exponential increases in the power of information and communication technologies, and rapidly declining costs for measures of computing power. In addition to widening the use of existing technologies through lowered costs, the increasing power enables new applications that were not technically feasible before. Both the reduced costs and the increased capabilities have favorable implications for telework.

Recent progress in communications technologies has positive implications for telecommuting. In the year 2003, digital subscriber line (DSL) or cable modem connections are available throughout most of the United States. For approximately $50 per month, telecommuters can access organizational networks and resources over the Internet at transmission rates of several hundred kilobits per second, without interfering with incoming or outgoing voice communications. This bandwidth is sufficient to enable real-time audio and video communications. With more than 3 billion Web pages indexed through the Google search engine in 2003, the telecommuting knowledge worker also has access to a tremendous wealth of information resources from all over the world.

The current situation is in is in sharp contrast to what was available just 10 years before in the early 1990s. At that time, telecommuters were limited to dial-up connections typically operating at 14.4 kilobits or less per second. The telephone costs for dialing in to corporate networks were so high that continuous online communications were impractical. Even with minimal dial-ins to remote computers, the potential interference with other calls often made it necessary to lease an additional phone for voice communications. Few telecommuters had heard of the Internet, and the World Wide Web was in its earliest stages, primarily being used by academic researchers.

The technological future promises to be even more favorable for telecommuting. Increasing Internet bandwidth will make it possible for off-site workers to reduce the impact of two problems—limited participation in meetings and reduced access to documents—that have been a significant impediment to telecommuting in the past. High-resolution audio and video teleconferencing will enable telecommuters to "attend" on-site meetings, at minimal cost, with a presence that is not markedly inferior to that of people who are physically in the conference rooms. This higher bandwidth will provide inexpensive real-time access to organizational document management systems. These increasingly popular systems store documents in a digital form that can be transmitted over internal networks and the Internet. In addition to reducing costs, these systems greatly facilitate "bringing the work to the worker," making it possible for employees to effectively handle a larger proportion of their tasks from remote locations.

Nonetheless, it remains to be seen whether these increased technological capabilities, by themselves, will cause telecommuting to become as widespread as Toffler predicted. It may be that there are other factors at work that will have a major impact on telecommuting participation and rates.

THEORETICAL CONSIDERATIONS

We are all aware of the tremendous advances in supporting technologies. Most of us are also only too familiar with the traffic congestion in major urban centers, which could make classic telecommuting more attractive in most developed countries. Many other considerations (e.g., high fuel taxes outside of the United States, environmental issues including global warming, and the political instability of many of the major oil producing nations) strongly favor increased usage of classic telecommuting. Figure 2 graphically illustrates this perspective, in which explicit factors have the greatest impact on telecommuting usage.

When expressed as a percent of total working hours, however, the actual substitution of telecommunications for vehicle trips is still low, not more than 3% at the start of the 21st century. This low usage represents a paradox, in view of the very favorable trends over the three decades in which the concept has been seriously discussed. In this same period, there have also been numerous implementation projects and many research studies. The continuing low usage despite all this suggests that there may be other, less obvious factors that are retarding the substitution of telecommunications for physical travel. Figure 3 illustrates the concept that telecommuting usage may be more affected by implicit factors, which have a greater impact than the explicit factors illustrated in Figure 2.

What kind of implicit factors could account for the low usage of telecommuting, despite the favorable explicit considerations? For possible explanations, I look to two theories: agency theory from the field of economics and (neo)institutional theory from sociology.

Figure 2: Explicit advantages and disadvantages model.

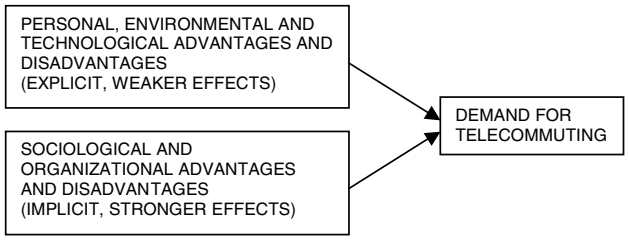


Figure 3: Explicit and implicit advantages and disadvantages model.

Agency Theory

This theory views relationships between managers and employees (and other economic relationships) as implicit contracts. In such relationships, an employee (agent) can be compensated either on the basis of behavior (what he or she is doing) or on outcomes (results, e.g., commissions on sales). In a typical office environment, where employees are physically present every working day, it is relatively easy to monitor behavior.

The more days per week an employee telecommutes, the more difficult it is to monitor behavior, and it becomes increasingly necessary to evaluate outcomes. For employees who are doing anything other than very routine processing, however, it is often difficult to measure outcomes. Typically it requires more managerial time (an expense item) to evaluate outcomes than monitor performance. Thus agency theory provides a plausible explanation for the often-noted reluctance of managers to allow telecommuting, and also the prevalence of 1-day-per-week classic telecommuting arrangements.

On the other hand, there have been numerous successful projects, in which organizations have generated large real estate savings by having salespersons and consultants work largely away from organizational offices. Although it may be more appropriate to identify these employees as mobile workers, the continuing usage of these arrangements (in contrast to the many classic telecommuting implementations that have not persisted) is exactly what agency theory would predict. Because such employees typically work on commissions or generate billings, they are already operating under outcome-based contracts. Therefore, the shift away from organizational offices requires relatively minor adjustments by their managers. (For more information on agency theory, see Eisenhardt, 1989.)

Institutional Theory

Sociologists consider the business office to be an institution: something that is considered by the larger society as the norm, the "right" way to do knowledge work. These norms transcend distances and cultural differences; for example, business offices in the United States, Asia, and Africa are more similar or homogenous than the underlying cultures and political systems.

Although office work is highly institutionalized, classic telecommuting has not yet reached that status. (As an indicator, although business offices often form the setting for scenes in movies or television shows, telecommuting workers are rarely seen.) It has a certain air of illegitimacy: other employees may view a person who is telecommuting as "not really working" while away from the office (e.g., Scott Adams' portrayals of telecommuters in the *Dilbert* comic strip often reflect this perspective). Neighbors may wonder if the telecommuter has lost his job. Telecommuting consultants frequently recommend increased managerial planning with and supervision of telecommuters, but from the perspective of institutional theory, this raises additional questions about the legitimacy of this mode of work in the minds of both employees and managers.

Institutional theory thus provides an alternative explanation for the low rate of telecommuting. Most people don't telecommute at all, or very much, because it is not as socially acceptable as traditional office work. This theory also accounts for the frequent association of telecommuting and health issues, for example, doing some work at home on days when an employee is ill or has a medical appointment. Telecommuting becomes more legitimate if there is an "excuse" for it, a connection with something that is recognized as an acceptable reason for being away from the office. (The seminal work on neo-institutional theory is Meyer and Rowan, 1977.)

RECOMMENDATIONS FOR STAKEHOLDERS

One of the impediments to increased classic telecommuting is that there are a number of types of stakeholders, and what is good for one stakeholder is not necessarily a benefit for all the others. Therefore recommendations must be addressed to the various stakeholders, but also must be harmonious with the needs and expectations of the others. It is appropriate to start with organizations, because organizational policies and attitudes are usually the most important factor in determining whether specific employees can telecommute and how much telework occurs.

Organizations

Employers need to consider telecommuting from a holistic perspective, so that they can deal with it based on its impacts on the overall performance of the whole organization, rather than just looking at how it affects the telecommuters. In many cases, this perspective will lead to organizational policies that are not directed specifically toward telework but which, as a side effect, will lead to modest increases in classic telecommuting.

Organizations need to have contingency plans and organizational capabilities that make it possible to function effectively in the event of a disaster. The more employees who have computer equipment and high-speed connections at home and who are comfortable using these capabilities, the better prepared they are for disasters. These capabilities also enable employees to work more effectively at home after and before regular working hours at the office, contributing to increased organizational productivity. And finally, this kind of infrastructure at home can lead to increases in classic telecommuting.

Therefore, organizational policies that make it easier or less expensive (or both) to acquire quality hardware, software, and telecommunications services can be very

cost-effective. Organizations can often provide such encouragement at relatively low cost by working out volume buying arrangements for hardware, software, and telecommunications services.

Various kinds of training can lead to improved organizational performance in general and also have positive implications for telecommuting. Training for specific kinds of software used at the office, for example, integrated e-mail and calendaring systems, will also lead to more efficient usage of the same software on home computers. Training in time management has obvious implications for better employee performance both at the office and while telecommuting. Training in management skills can make managers more effective in handling employees at the office, mobile workers in the field, and employees who do classic telecommuting 1 day or more per week.

Telecommuting arrangements sometimes make it possible for organizations to retain key employees who would otherwise be lost because either the employee or the organization relocates. Telework arrangements may make it possible for an organization to access employees with skills (e.g., computer programming) that might not otherwise be available. These situations generally are unique enough to be handled on a case-by-case basis, rather than through formal organizational policies.

Employees

Telecommuting offers substantial personal benefits, even if it is only done for an average of 1 day a week. Some of these benefits—vehicle costs and reduction in commuting stress—are proportional to the length of the commute. For example, at the standard U.S. rate of 36 cents per mile in 2003, telecommuting 48 days per year instead of driving back and forth to a work location that is 15 miles away would amount to a savings of more than $500 (U.S. dollars, undiminished by income taxes) per year. For some employees, the reduction in commuting stress (White & Rotton, 1998) can be important, especially when bad weather or other factors make the drive worse than usual. In some relatively rare situations, full-time telecommuting makes it possible for an employee to maintain a desirable position within an organization after either the employee or the organization relocates.

These personal benefits need to be placed within the context of the employee's standing in the organization and the needs of the employer, however. If an employee possesses skills that would be difficult to replace, based on specialization or on experience with the organization, it may be possible for him or her to work out a favorable telecommuting arrangement. Similarly, if the characteristics of the job are such that telecommuting makes a lot of sense for the organization, it may not be difficult to have a telecommuting arrangement. For commission sales, and telephone customer service representatives outside regular working hours, it may be more cost-effective for the organization and more productive for employees to work most of the time out of their homes.

If the employee does not have that kind of a bargaining position or job description, the emphasis would necessarily shift toward the perspectives of (a) how telecommuting can improve the individual's performance and (b) how it can contribute to the overall effectiveness of the organization. Although their organizations may not mention this explicitly, it would be a good idea for such telecommuters to reinvest their commuting time savings into extra work on organizational projects. In a sense, telecommuting is a privilege, so it would be appropriate for telecommuters to demonstrate that their organizations are getting something of value in exchange for any inconveniences resulting from their physical unavailability. Although there is little empirical backing for the claims of increased productivity, telecommuters would be wise to put in the extra time to fulfill the expectations that these repeated claims may have generated in the minds of their managers and within their organizations.

A high level of flexibility in scheduling telecommuting days is important to avoid inconveniences to other employees. Employees who telecommute 1 day per week should, as much as possible, avoid telecommuting by necessity on a certain day of every week. Whenever there is a conflict between a scheduled telecommuting day and organizational requirements, the employee should try to move the telecommuting to another day or, if that is not possible, forego telecommuting for that week.

Telecommuting should also be based on identifying tasks that can be accomplished significantly better when working away from the office. If a task can be done equally well at the office or at home, in most cases it should be handled at the office to avoid inconveniences to other employees. Based on this criterion, there may be weeks where telecommuting is not appropriate at all and other weeks where more telecommuting than usual can be justified.

Telecommuting Equipment, Software, and Other Considerations

The requirements for hardware, software, and furnishings will vary depending on the type of work done while telecommuting, the frequency of telecommuting, the type of software and systems being used within the organization, and the telecommuter's personal preferences and technical capabilities. Cable or DSL connections are desirable for almost all telecommuters, because of the extra bandwidth and because they do not interfere with use of a single telephone line for voice communications.

The same business productivity software (word processor, spreadsheet, etc.) should be used at home as at work to avoid time-consuming compatibly and conversion problems. For software categories that are less vulnerable to interoperability issues (e.g., e-mail systems, web browsers, computer program development environments), technically savvy telecommuters would do well to use software that differs from organizational standards if it makes them significantly more productive without creating problems for coworkers.

Regulatory Authorities

At the public policy level, it is important to be aware that telecommuting participation and usage trends have not and probably will not result in noticeable reductions in the need for transportation infrastructure (see Mokhtarian, 1998). At the start of the 21st century, the total traffic volume reduction from classic telecommuting throughout

the United States is not more than 3%, which is less than the annual traffic volume growth per year in some major urban areas in the country.

Responding to their employees' concerns and the impacts on local labor markets, businesses have been effective in opposing regulations, such as parking space restrictions, that indirectly encourage telecommuting by making it more difficult for their employees to drive to work. It is generally not worth the effort to enact stringent regulatory disincentives for alternatives to telecommuting.

In an era of tight budgets and increasing demands for government services, it is also difficult to justify governmental expenditures (direct spending or tax incentives) that would encourage telecommuting, because in most cases the benefits will not justify the costs. The best course of action at the public policy level would be to let organizations and their employees work this issue out on their own with minimal government intervention. This laissez faire stance would include avoiding policies that might have the side effect of discouraging telecommuting. As an example of unfavorable policies to avoid, the Occupational Health and Safety Administration proposed standards that would require employers to inspect conditions in employees' homes if they used their home computers to do any job-related work (Joyce, 2000), but this proposal was dropped after encountering strong opposition.

CONCLUSION

After 30 years of research and experience with the concept, it is evident that classic telecommuting is not, and is unlikely to ever become, a panacea for problems in the areas of transportation, the environment, or oil-related foreign policy. Commonsense analyses of findings of telecommuting research also indicate that it does not generate sustained large improvements in organizational productivity that are noticeable at the bottom line.

Classic telecommuting is and will continue to be one option within organizational human resources policies. Even though it has implications for all of the following, by itself it is not and will not be as important a human resources issue as recruiting, compensation, or training. In most organizations, telecommuting will continue to be implemented on a case-by-case basis in response to the needs of specific organizational units and individual employees who may not be typical of the organization.

The continuing development of the Internet and other information technologies should contribute to increases in the proportion of work days where employees can substitute telecommunications for commuting without creating organizational problems. The huge increases in information technology capabilities and great reductions in information and communications technology costs have had a limited impact on this substitution over the 30 past years, however. It is questionable whether such technology trends, in the absence of other, more compelling considerations, are going to have a substantially larger impact on classic telecommuting over the next 30 years.

On the other hand, other forms of telework will grow rapidly. This growth should to continue until it approaches saturation of the employees whose job functions require them to work outside of organizational facilities at least part of the time. One reason for the differing outcomes is that, in contrast to the effects of classic telecommuting, this form of telework actually increases the amount of contact between remote workers and other members of the organization.

GLOSSARY

After-hours telecommuting Using information and communications technologies to perform organizational work after (or before) regular working hours.

Classic telecommuting Substituting information and communications technologies for physical commuting to a work location.

Free-address workspaces Unassigned organizational workstations available to employees on a first-come, first-served basis.

Full-time telecommuting Classic telecommuting on (usually) every day of the work week.

Home workers Persons who use their homes as the base for doing most or all of their paid work.

Hoteling Providing organizational workspaces that are available on a reservation basis.

Mobile worker An employee who primarily works outside of organizational offices and uses information and communications technologies for communications with supervisors, for scheduling and routing, and so on.

Moore's Law (as commonly understood) The doubling of computer processing power every 18 to 24 months, without increases in costs.

Part-day telecommuting Telecommuting for part of a work day and working at organizational offices for the remainder.

Telecommuting center (sometimes called a satellite center) An office facility between employees homes and their regular workplaces, where they can work instead of driving all the way to their organizational locations.

Telecommuting frequency How often (e.g., how many days per week) classic telecommuters substitute telecommunications for physical travel to work.

Telecommuting participation The percentage of employees (in an organization, region, or throughout a national economy) who engage in classic telecommuting.

Telework Work at a distance from organizational or other facilities.

Virtual office A location or workplace that enables a person, by means of information and telecommunication technologies, to work together with others who are not physically nearby.

Virtual organization A (usually short-term) venture composed of components and personnel from different organizations that may be widely separated geographically, and that requires or is enabled by extensive use of telecommunications capabilities for coordination and communications.

CROSS REFERENCES

See *GroupWare; Internet Literacy; Virtual Enterprises; Virtual Teams.*

REFERENCES

Bailey, D. E., & Kurland, N. B. (2002). A review of telework research: Findings, new directions, and lessons for the study of modern work. *Journal of Organizational Behavior, 23*, 383–400.

Christensen, K. (1988). *Women and home-based work: The unspoken contract*. New York: Holt.

Eisenhardt, K. M. (1989). Agency theory: An assessment and review. *Academy of Management Review, 14*, 57–74.

Forster, E. M. (1909). The machine stops. *Oxford and Cambridge Review.* Retrieved March 13, 2003, from http://www.plexus.org/forster.html

Harkness, R. C. (1973). *Communications substitutes for travel: A preliminary assessment of their potential for reducing urban transportation costs by altering office location patterns*. Unpublished doctoral dissertation, University of Washington, Seattle.

Huws, U., Korte, W., & Robinson, S. (1990). *Telework: Toward the elusive office*. New York: Wiley.

"It's rush hour for telecommuting." (1984, January 23). *Business Week,* 99, 102.

Joyce, A. (2000, January 5). Reversal sought on OSHA telecommuting rule. *The Washington Post*, p. E01.

Kidder, T. (1981). *The soul of a new machine* (1st ed.). Boston: Little, Brown.

Kraut, R. (1989). Telecommuting: The trade-offs of home work. *Journal of Communication, 39,* 19–47.

Memmott, F. W., III. (1963). The substitutability of communications for transportation. *Traffic Engineering, 33,* 20–25.

Meyer, J. W., & Rowan, B. (1977). Institutionalized organizations: Formal structure as myth and ceremony. *American Journal of Sociology, 83,* 340–363.

Mokhtarian, P. L. (1996). The information highway: Just because we're on it doesn't mean we know where we're going. *World Transport Policy and Practice, 2(*1/2*),* 24–28. Retrieved March 13, 2003, from http://www.its.ucdavis.edu/telecom/nrp12.html

Mokhtarian, P. L. (1998). A synthetic approach to estimating the impacts of telecommuting on travel. *Urban Studies, 35,* 215–241.

Nilles, J. M. (1994). *Making telecommuting happen: A guide for telemanagers and telecommuters*. New York: Van Nostrand Reinhold.

Nilles, J. M., Carlson, F. R., Jr., Gray, P., & Hanneman, G. J. (1976). *The telecommunications–transportation tradeoff*. New York: Wiley.

Olson, M. H. (1989). Organizational barriers to professional telework. In E. Boris & C. R. Daniels (Eds.), *Homework: Historical and contemporary perspectives on paid labor at home* (pp. 126–134). Chicago: University of Illinois Press.

Prusak, L., & Cohen, D. (2001). *In good company: How social capital makes organizations work.* Boston: Harvard Business School.

Toffler, A. (1980). *The third wave*. New York: William Morrow.

Wells, H. G. (1899). *When the sleeper wakes*. Retrieved March 13, 2003, from http://www.neponset.com/books/sleeper/sleep00.htm

Westfall, R. D. (1997). Does telecommuting really increase productivity? Fifteen rival hypotheses. *Proceedings of the AIS Americas Conference.* Retrieved March 13, 2003, from http://www.cyberg8t.com/westfalr/prdctvty.html

White, S. M., & Rotton, J. (1998). Type of commute, behavioral aftereffects, and cardiovascular activity: A field experiment. *Environment and Behavior, 30*, 763–780.

Wiener, N. (1950). *The human use of human beings: Cybernetics and society*. Boston: Houghton Mifflin.

Trademark Law

Ray Everett-Church, *ePrivacy Group, Inc.*

INTRODUCTION

Trademark law is a fascinating subject for many people, in part because most everybody in our society understands and appreciates the power of popular trademarks, such as Lexus, Pokemon, Safeway, and Yahoo! Trademarks are such an integral part of our language and culture that we all have a vested interest in their protection. Because trademarks are all about meaning, trademark disputes are a kind of spectator sport: They involve popular cultural icons and turn on questions such as whether the average person is likely to be confused if a trademark is used improperly. So in many respects, everybody gets to have an opinion on trademark issues, and that opinion more often than not counts for something in the final calculus of trademark disputes.

In this chapter, I discuss the fundamental ideas that underlie the protection of trademarks and look at ways in which trademarks can be infringed and protected. Once that groundwork has been laid, I then look at how these fundamentals have been to be applied to the unique and significant disputes that have arisen in the Internet context. In many respects, trademark law has been turned upside down the Internet, so I describe how the principles of trademark law are being applied in today's Internet-oriented business environment.

TRADEMARK DEFINED

Merriam-Webster's Dictionary of Law defines *trademark* as "a mark that is used by a manufacturer or merchant to identify the origin or ownership of goods and to distinguish them from others and the use of which is protected by law."

In practice, a trademark is any word (Sun), name (Calvin Klein), symbol (golden arches), device (the Energizer Bunny), slogan ("Fly the Friendly Skies"), package design (Coca-Cola bottle), colors (FedEx purple and orange), sounds (the five-tone Intel Corporation sound), or any combination thereof that identifies and distinguishes a specific product or service from others in the marketplace.

As trademark law has evolved, the field has become an important subset of the larger category known as *intellectual property* law. As the name implies, intellectual property law (which also includes patent, copyright, and trade secret law) treats these rights as a kind of property right, protecting the rights of owners to exploit the property for their own benefit while prohibiting unauthorized use by others. Unlike real estate or personal property law, intellectual property law concerns ownership of intangible things such as ideas, words, and meanings, rather than physical things.

The legal protections afforded by trademark law also extend to the related concepts of *service marks* and *trade dress*. Service marks differ from trademarks in that they are marks used to identify a particular service or to distinguish the provider of a service, rather than a tangible product. For example, the name of a consulting firm, or the name of a proprietary analytical process used by that consulting firm, might be more properly identified as a service mark. Trade dress is the overall image of a product, composed of the nonfunctional elements of its design, packaging, or labeling. This could include specific colors or color combinations, a distinctive package shape, or specific symbols or design elements.

Many people confuse trademark with copyright. Copyright is a person's exclusive right to benefit from the reproduction or adaptation of an original work of authorship, such as a literary, artistic, or musical work. Trademark differs from copyright in that trademark law does not prohibit the reproduction or adaptation of the creative products of an author. Rather trademark law seeks to prevent confusion over words or other characteristics used to identify uniquely the source or quality of a product or service. For example, Paul Simon's 1973 song "Kodachrome" refers to a trademark owned by Eastman Kodak Company even though Simon holds the copyright on his work.

The relative strength of a particular trademark depends upon where it falls within a range of five categories: fanciful, arbitrary, suggestive, descriptive, or generic. The greatest protection comes for fanciful marks consisting of invented words such as Xerox, Kodak, or TiVo. The next strongest protection comes for arbitrary marks, which are commonplace words used in a manner that is unrelated to their dictionary meaning, such as Apple for computers or Shell for gasoline. Suggestive marks are familiar words or phrases that are used in an inventive way to "suggest" what their product or service really consists of, such as Home Box Office for a movie channel or Mail Boxes Etc. for a postal services franchise. The least protection comes for descriptive marks that do little more than describe the characteristics or contents of the product, such as the publication Automotive Industry News, or Cellphone Center for a cellular telephone retailer. Finally, generic names, which merely state what the product or service is, cannot function as trademarks. Some marks, such as aspirin, linoleum, escalator, or nylon, were once trademarks but became generic because the trademark holder failed to police unauthorized use.

FEDERAL TRADEMARK LAW

For trademarks used in interstate commerce, U.S. law provides protection under the Trademark Act of 1946, known more commonly as the *Lanham Act*. The Lanham Act also created a registration process for trademarks and legal and procedural incentives for trademarks to be registered with the U.S. Patent and Trademark Office (USPTO) (USPTO, n.d.). Many states within the United States also afford trademarks protections under their state's laws.

The Lanham Act provides a functional definition of what is eligible to be registered as a trademark. Potentially anything can be registered as a trademark if it functions among consumers to distinguish a specific product from other products in the marketplace. The Lanham Act does, however, prohibit certain marks, including anything immoral, deceptive, scandalous, disparaging toward an institution or national symbol; anything that falsely suggests a connection to a person, that consists of a flag other governmental insignia, or that uses a name or portrait of a deceased U.S. president during the life of his widow without her consent; and numerous other limitations. Once registered, a mark may also be cancelled if it has become generic, has been abandoned, was obtained through fraud, or is otherwise prohibited by the aforementioned conditions.

One important limitation on registration comes when the mark is part of the product's functionality. An aspect of the product may meet the definition of a trademark and may even be recognized as a trademark by consumers, but it cannot be registered if it is essentially a functional aspect of the product. For example, a company might sell a computer monitor with a unique shape that is immediately recognizable to the public and distinguishes the monitor from the products of competitors. Under the functional definition of a trademark, the unique shape may be registered. If the shape is actually a functional aspect of the product, for example, if the shape is responsible for improved resolution, the shape cannot be registered as a trademark. (In such a case, however, the manufacturer may be able to seek patent protection for the unique design.)

The USPTO registers trademarks and service marks that are used in interstate commerce. Trademarks need not be registered for an owner to enforce his or her rights in court; however, federal registration provides numerous legal benefits to a trademark owner at a reasonable expense. For example, once a mark is registered, the registration establishes the validity the registrant's claims of ownership and places the world on constructive notice that the owner has exclusive rights to use the mark in commerce. If the holder of a registered trademark establishes infringement under the Lanham Act, the holder not only can enjoin any misuse of a mark but also can recover statutory damages and, in some cases, attorneys' fees.

Federal registration on the principal register gives nationwide protection from infringement, whereas common law protects the mark only in the specific geographical area in which the mark is used in commerce; and state law protects the mark only within the state where the mark is registered. Thus, another benefit of federal registration is the establishment of rights across a larger geographical area than under common law and state law.

In fact, the scope of protection for a federally registered mark is usually broader than under common law or state law. For example, under common law and many state trademark registration statutes, trademark protections may be restricted to those specific products or services for which the mark has explicitly been used, whereas federal law allows a mark to be protected even when used in conjunction with a wider array of related products or services, such as a family of services offered under an umbrella trademark. Finally, the Lanham Act provides legal remedies that go beyond those available at common law including, for example, treble damages against a "willful" infringer, as well as reimbursement of attorney fees in exceptional cases. It is important to note, however, that unlike many areas in which federal law supercedes state law, state and federal trademark law often co-exist well and aggrieved parties can frequently bring legal actions using both state and federal law to equal effect.

The Lanham Act's protections flow equally to trademarks and service marks. Trade dress is also protected by the Lanham Act, provided it is not a functional part of the product and is distinctive, has acquired secondary meaning as being uniquely associated with the product, and there is a likelihood of confusion on the part of the consumer if a competing product were to possess similar trade dress.

Trademark Registration

The USPTO maintains two types of trademark registries, the Principal Register and the Supplemental Register. The Principal Register is where a "registered trademark" is registered. There are three ways a trademark or service mark may be registered with the USPTO. The first method, called an "in use" application, is for an applicant who is already using a mark in commerce. The second method is an "intent to use" application, for marks that are not yet in use but that the applicant is preparing to use. The third

method is based on certain international agreements, by which applicants outside the United States can file an application based on applications or registrations in another country.

The Supplemental Register is where marks that are descriptive in nature but have not yet established secondary meaning are maintained. Marks on the Supplemental Register can use the ® symbol, and if the mark is continuously used and unchallenged for 5 years, the holder may file another application and claim such use presumptively establishes secondary meaning under Section 2(f) of the Lanham Act, and thereby move the mark onto the Principal Register.

The registration process itself is relatively straightforward. The application documents must be filed by the owner of the mark, usually through the services of an attorney concentrating in trademark law. (For brevity this chapter focuses only on an "in use" application and does not discuss further the "intent to use" application.) The application contains information about the individual or corporation that owns the mark, an exact representation of the mark (in text or in image form), as well as several specimens of the mark in actual use, information about the date of first use and date of first use in commerce of the mark, a description of the goods or services used in conjunction with the mark, and the "classification" of the goods or services according to a standardized list of 42 predefined classifications. Some goods and services may be registered in multiple classes, with the application fees increasing accordingly.

Once received, the USPTO makes an initial review of the application to determine if the application contains all the information necessary to be considered "filed." If the application is complete, a "filing date" is issued along with a serial number and sent to the applicant. Several months after filing, an examiner at the USPTO reviews the application in more detail, researches the information provided, and makes a determination as to whether the mark should be registered. If it cannot be registered, the examiner will issue a notice called an "office action," which explains the grounds for refusal, including any deficiencies in the application itself. In some cases, only minor adjustments might be necessary to permit registration, and sometimes the application can be corrected over the phone. Applicants have six months to respond to an office action before the application is considered abandoned. If the applicant cannot overcome the examiner's objections, a final office action is issued, at which point the applicant may appeal to the Trademark Trial and Appeal Board. Should the applicant be unsuccessful there, that decision may be appealed to federal court.

If there are no objections to the application, or the applicant overcomes the objections, the examiner will approve the mark for publication in the *Official Gazette,* a weekly publication of the USPTO. The applicant is notified of the date of publication through an official Notice of Publication.

Anyone who believes the registration harms him or her or that it is otherwise in violation of the Lanham Act has 30 days from publication to file an opposition to the registration or seek an extension of time to do so. At this point, the administrative proceeding is *inter partes* (meaning between two parties, in contrast to an *ex parte* proceeding before the examiner only) and is known as an "opposition." The opposition proceeding determines the validity of the objections. If no objection is received, the mark will be registered. After the registration is issued, anyone who believes himself or herself to have been harmed by the registration may begin a "cancellation proceeding," which is similar to an opposition proceeding except that it takes place after registration. In an opposition, the applicant bears the ultimate burden of establishing registerability; in a cancellation, the party seeking the cancellation bears the burden of proving the registration was improvidently issued. Opposition and cancellation proceedings are held in a formal, trial-like hearing before the Trademark Trial and Appeal Board, a division of the USPTO.

A mark will be registered only after it has been published and the opposition period has expired. Once registered, federal trademark registrations run for 10 years, with renewal terms lasting 10 years. Between the fifth and sixth year, however, the registrant must file an affidavit to confirm the mark is still in use. If that affidavit is not filed, the registration is cancelled. Thus, a trademark must remain in use or its registration may be cancelled, but if a mark is in continual use and that use is properly demonstrated as required by law, the registration could remain effective forever.

After five years, the owner of a registered mark may request that the mark be deemed "incontestable." Under the Lanham Act, incontestability means that certain legal avenues of challenging the mark-such as a claim that the mark is not distinctive, lacks secondary meaning, is confusingly similar to another mark, or the mark is purely functional—are no longer available. The term "incontestable" is somewhat misleading in that there remain certain circumstances in which the mark may be challenged and have the registration cancelled, such as an assertion that the mark was improperly registered in the first instance.

The Differences Between ®, ™, and ℠

Once registered, the registration symbol, ®, may be used. It is considered trademark misuse to display the registration symbol at any point before the USPTO issues the final registration notice to the applicant. In contrast, anyone who wishes to claim rights in a mark may use the ™(trademark) or ℠ (service mark) designation along side the mark. Use of ™ or ℠ alerts the public to the claim of ownership and exclusive use. It is not necessary to have a registration, or even a pending application, to make use of these designations, and consequently the claim may or may not have any validity. In short, use of ™ or ℠ tells the world that the party using the trademark is prepared to put up a fight.

STATE STATUTES AND COMMON LAW

Many states also have trademark registration statutes that allow registration of marks used in intrastate commerce, using procedures that function similarly to the USPTO process defined by the Lanham Act. In addition, all states

protect unregistered trademarks under some combination of state statute and common law. For the sake of brevity, this section does not detail trademark protections in all 50 states.

In most states, the common law recognizes ownership of a trademark. Ownership under common law is most often established by demonstrating when the mark was first used in commerce, but unlike federal law, common law protections extend only to those areas or markets in which the mark is actually used. In contrast, federal registration of a trademark gives a basis under federal law for a suit for infringement, in addition to any common law claims that might be available. Although it is possible to protect one's rights using only common law or state statutory protections, the benefits that flow from federal registration make it highly desirable.

INFRINGEMENT AND DILUTION

There are two main rights that trademark owners will assert: *infringement* and *dilution*. Infringement of a trademark usually involves the use of the mark in a way that is so similar to the owner's usage that the average purchaser will likely be deceived, will mistake the infringing goods for the original, or will likely experience confusion. Dilution is a lessening of the value of a trademark caused by an unauthorized use of the mark, regardless of whether any actual confusion, deception, or mistake occurred.

Infringement

Under the Lanham Act, the standard for determining whether a mark is infringing is whether there is a "likelihood of confusion" over the mark in a particular usage context. More specifically, infringement comes when a consumer is likely to be confused over the source, sponsorship, or approval of the goods bearing the mark.

In deciding whether consumers are likely to be confused, courts have previously looked at a number of factors, including the following:

- Similarity between the two marks (such as any visual, phonetic, or contextual similarities),
- Similarity of the goods or services being offered,
- Proximity of the goods in a typical retail setting,
- Strength of the plaintiff's mark as exemplified by how well known the mark is to the public at large,
- Evidence of any actual confusion by consumers,
- Evidence of the defendant's intent in using the mark,
- Likely level of care employed by consumer in the purchase of that type of product, and
- Likely growth or expansion of the product lines.

Of these eight factors, the first two are arguably the most important. For example, using an identical mark on an identical product is a clear case of infringement, such as a company other than Ford manufacturing a midsized automobile and calling it a Taurus. Similarly, calling the vehicle a Taurius would run into problems. (This use of similarly spelled names is of particular concern in the Internet domain name context, which will be discussed in a later section.)

Mere similarity is not always determinative of infringement. For example, it is possible to find Delta Faucets just a few aisles away from Delta power tools at your local home improvement store. Although made by different companies, the similarity in trademark does not constitute infringement because consumers are not likely to mistake a belt sander for a shower head.

Dilution

To clarify further the distinction between normal trademark infringement and dilution, in 1995 Congress amended the Lanham Act by passing the Federal Trademark Dilution Act (FTDA). This legislation expanded protections granted to famous and distinctive trademarks under the Lanham Act. Unlike infringement, dilution does not require evidence of a likelihood of confusion. Instead, the plaintiff must demonstrate that their mark is "famous," that it is being used in commerce by another party, and that the use causes the dilution of the "distinctive quality" of the mark.

The FTDA says that in determining whether a mark is "famous," a court may look at factors including the length of time a mark has been used, how widely and in what geographic areas it as been advertised, how recognizable the mark is to the public, and other factors. Highly distinctive, long-used, and well-known marks, such as Coca-Cola or Kodak, are examples of famous marks. Once a plaintiff establishes the fame of a mark, the owner can seek an injunction against further use of the mark in a manner that dilutes the distinctive qualities of that mark.

There are two types of dilution of a mark: *blurring* and *tarnishment*.

Blurring

Blurring is the weakening of a mark through its identification with dissimilar goods. For example, marketing Kleenex brand refrigerators would not likely confuse someone looking for bathroom tissue and cause them to purchase a refrigerator accidentally; however, the use of the trademark would dilute the mark's distinctiveness as representing personal care paper products.

Tarnishment

Tarnishment is the use of a mark in an unflattering light, through associating it with either inferior or distasteful products. For example, in the case *Toys 'R' Us v. Akkaoui,* (1996), the toy retailer brought a successful tarnishment claim against a pornographic Web site "adultsrus.com."

Other Trademark Claims

Although dilution claims and infringement claims based on likelihood of confusion are the two most common trademark-related causes of action, there are a number of other bases for bringing suits. Many states have enacted unfair competition laws that prohibit a range of activities known as *passing off, contributory passing off, reverse passing off,* and *misappropriation.*

Passing Off

Passing off occurs when a defendant attempts to "pass off" its product as if it were the mark owner's product. For example, affixing a Dell nameplate to computers actually made in someone's basement would constitute passing off.

Contributory Passing Off

Contributory passing off occurs when a defendant induces a retailer to pass off a product. For example, bribing a computer store to sell computers with a fake Dell nameplate would be contributory passing off.

Reverse Passing Off

Reverse passing off takes place when someone tries to market someone else's product as their own. If a computer store purchased Dell computers, replaced the nameplate with its own store brand nameplate, and attempted to sell the computers, it would have engaged in reverse passing off.

Misappropriation

Misappropriation, a privacy-related tort, is traditionally defined as using the name or likeness of someone for an unauthorized purpose, such as claiming a commercial endorsement by publishing someone's image (or even that of a look-alike impersonator) in an advertisement. In the trademark context, using a mark without authorization can violate federal and state law prohibitions on certain unfair trade practices, including the unauthorized use of marks in inappropriate ways.

Parody and Fair Use

Aside from challenging the validity of a trademark claim or attacking the elements of the infringement claim, defendants in trademark infringement or dilution cases can also claim two affirmative defenses: parody and fair use.

Parody

Certain uses of a trademark for purposes of humor, satire, or social commentary may be permissible if they are not closely tied to commercial use. The theory underlying the protection of parody is that artistic and social commentary are valuable contributions to the society, therefore some deference to the First Amendment's protection of these types of speech is in order, even when balance against the detriment to a trademark owner. The protections vary, however. For example, in the highly amusing case of *Hormel Foods Corp. v. Jim Henson Productions* (1996), the use of a piglike character named "Spa'am" in a Muppet movie was found not to violate Hormel's rights in the trademark "SPAM." In *Coca-Cola Co. v. Gemini Rising, Inc.* (1972), however, the printing of posters with a stylized slogan and logo reading "Enjoy Cocaine" were found to violate the rights of Coca-Cola in the stylized slogan and logo "Enjoy Coca-Cola."

Fair Use

Fair use occurs when the public benefit of allowing the use is perceived to override any perceived harm to the trademark owner. For example, in the case *Zatarains, Inc. v. Oak Grove Smokehouse, Inc.* (1983), the defendant's use of "fish fry" to describe a batter coating for fish was not an infringement of the plaintiff's mark "Fish-Fri." The court held that fair use prevents a trademark owner from monopolizing a descriptive word or phrase to the exclusion of other parties that seek merely to describe their goods accurately. The defense of fair use is only available, however, when the mark at issue is descriptive, and then only where the descriptive term is used descriptively. Federal trademark statute also contains a right to fair use limited to usage in comparative advertising.

POLICING TRADEMARK ON THE INTERNET

Along with the tremendous growth in the usage of the Internet for both commercial and personal use, there has been a similar expansion in the number of trademark-related disputes involving the Internet. In a later section, I discuss the complex legal issues arising from trademark disputes over Internet domain names. First, however, there are a number of trademark issues that arise just from the very nature of the Internet as a facilitator of ubiquitous information sharing and access.

Perhaps the most important reason behind the growing amount of trademark-related litigation is that uncovering instances of trademark violations can be as simple as typing your trademark into an Internet search engine. Just a decade ago, a trademark owner in Maine might have no idea that his trademark might be in use by someone in Oregon. With the ability to search the Internet, trademark owners are quickly able to perform searches that might have been impossible—or just impossibly costly—a few years ago.

The ability to discover trademark infringement so easily, both intentional and unintentional, has catapulted trademark law into one of the most active areas of litigation in the Internet arena. The nature of trademark law itself has also added to the litigation explosion. As noted earlier, failure to police a mark properly can result in it becoming generic, and thus unprotected. Therefore, the same ease with which a trademark owner might uncover infringement may require that a trademark owner keep policing the Internet routinely and bring enforcement actions: If an infringement is known—or could be discovered through basic due diligence—and goes unchallenged, the trademark owner could lose control of its mark.

The requirement of constant policing of trademarks has, however, caused the unfortunate side effect of a growing number of heavy-handed actions against inexperienced Web users, and still more enforcement actions that are brought in cases in which a finding of infringement or dilution is highly unlikely. In many of these cases, well-intentioned individuals have been bullied by corporations over trademarks appearing on personal Web pages. In some cases that have received significant media attention, sites created by fans of rock groups, automobiles, and movie stars have been threatened by the very entities that the sites were set up by their creators to honor.

Meta Tags

In recent years, several disputes have arisen over the use of trademarks in *meta tags* on Internet Web pages. Web

pages are coded using a type of programming language called hypertext markup language, or HTML. The codes that are embedded in HTML documents, called tags, tell the browser how to display the information contained on the page, such as when to display words in bold, when to change fonts or font sizes, how to align tables, or where to place images. Web page designers can also include meta tags, which are special tags that contain information about the contents of the Web page. Meta tags are used by search engines to find and rank pages so that more relevant search findings are displayed before less relevant ones.

In one of the first lawsuits over meta tags, *Oppedahl & Larson v. Advanced Concepts, et al.* (1997), a Colorado law firm discovered that the defendants had put the law partners' names, "oppedahl" and "larson," in meta tags on several Web pages. This was presumably done in hopes that searches for the respected law firm's name would gain more attention for the defendants' Web pages. Suing under both the Lanham Act and the Federal Trademark Dilution Act, as well as state and common-law unfair trade practice actions, the law firm won a permanent injunction against any further use of its names in meta tags on the defendants' Web sites. Since that case, a number of other disputes have tested the extent to which trademarks may be used in meta tags and have largely resulted in prohibitions against uses by entities seeking to enhance site traffic by using the marks of competitors.

One of the issues that has arisen in meta tag disputes is the concept of "initial interest confusion." Initial interest confusion occurs when the use of another's trademark is done so in a manner reasonably calculated to capture initial consumer attention, even though no actual sale is finally completed as a result of the confusion. The case of *Brookfield Communications, Inc., v. West Coast Entertainment Corp.* (1999) illustrates the issue. Brookfield operated a Web site, MovieBuff.com, containing a movie database. West Coast, a video retailer, used the term "moviebuff" in meta tags on its Web site. A court held that West Coast's use of term in meta tags led to "initial interest confusion," in which search engine users looking for MovieBuff.com's site might visit West Coast's site and stop looking for MovieBuff.com, even though there might never be any confusion over sponsorship of the two sites.

Not all cases in which meta tags were at issue have resulted in a ban on their use. For example, in the case of *Playboy Enterprises, Inc. v. Terri Welles* (2002), a model who had posed as a Playboy Playmate of the Month was permitted to use "Playboy" and "Playmate" as meta tags for her Web site.

Deep Linking

Fundamental to the functioning of Web pages on the Internet is the concept of a link. A link, short for hyperlink, is a tag coded within a Web page that turns a piece of text (or in some cases an image) into a pointer to another document or page. Clicking on that link will typically cause the browser to follow the link and open the new page. Although a simple link to the home page of a Web site will not typically run into trademark issues, some sites choose to create links to pages many levels down within a site. For example, instead of linking to the home page of a manufacturer, a Web site designer might choose to create a link that goes directly to a page displaying one of the manufacturer's products. This practice is called *deep linking*.

Some site owners object to deep linking because it allows visitors to quickly bypass other contents of a Web site, including advertisements, which they would normally see if they had to navigate step-by-step through the contents of a site. In several court cases, plaintiffs have charged that deep linking deprives them of the full benefits of having visitors explore their site and have argued a variety of copyright, trademarks, and unfair competition claims. Proponents of deep linking counter that deep links are no different from footnotes or bibliographies, permitting readers to jump quickly to precise information. There are few clear court decisions on the trademark implications of deep linking; however many of the suits have focused on evidence of a defendant's bad faith, such as any appearance that the deep linking is intended to take unfair advantage of the other site's content, which will cut strongly in favor of the plaintiff.

In a related issue, there have been numerous disputes over the practice of "framing" Internet content. Framing is a technique in which content from one site is displayed within a "frame" appearing on another unrelated site. The use of framing often makes it appear that the content is owned or otherwise presented by an entity other than its actual owner or authorized user. Most disputes regarding framing have centered on copyright implications of unauthorized framing of content; however, trademark issues also arise when there might be confusion as to the source of the content or its relationship to advertisements and other affiliations that might be suggested by the way in which the framed material appears.

DOMAIN NAMES

With the explosive growth of the Internet, both in its importance to global commerce and in the effect it has had on all aspects of our society, the importance of the domain names used on the Internet cannot be understated. The academic and noncommercial roots of the Internet caused many of its key functions, such as the domain name system, to be designed without some important safeguards. For example, domain names could then—and in many cases can still—be registered by anyone willing to pay the registration fee. In the early days of the Internet, this fact caused something of a "land grab" mentality in which speculators rushed to purchase the rights to domain names that were expected to become valuable. Indeed, the domain name WallStreet.com, registered for under $100, was reportedly sold for more than $1 million (Bicknell, 1999).

Unfortunately however, some speculators also rushed in and purchased domain names that were identical (or in some cases merely similar) to valuable brand names. These so-called cybersquatters sought to gain financially by occupying the "virtual" real estate of someone else's trademark translated into a domain name. Because a domain name has become such an important part of a company's marketing identity, trademark owners have been

forced to wage legal battles to retake control of their trademarks in cyberspace. Cybersquatting is discussed in detail in a later section, but it may be useful to look first at how domain names work and why they have become a trademark law battleground.

Domain Name System Basics

Generally speaking, each computer connected to the Internet requires a unique address, called an Internet protocol (IP) address, in order to distinguish it from all the other computers on the Internet. When computers communicate across the Internet, they use IP addresses to ensure that when a user on a particular computer requests data from another computer, the data gets delivered to the right place.

IP addresses are not friendly to human eyes. Looking something like "192.168.27.145," it was quickly determined that it would be easier to assign names to stand in for those numbers because many humans find it easier to remember names than to remember numbers. Thus, the designers of the early Internet developed the domain name system (DNS) to permit the reliable association of names with IP addresses. As a result, with the help of DNS, when users tell their Web browsers that they want to check out the latest news at CNN.com, it is able to direct the query to 64.236.16.116, which is one of the many Web servers that answer to the busy CNN.com domain name.

Domain names, and their underlying numbers, are controlled by the Internet Corporation for Assigned Names and Numbers (ICANN). ICANN controls not only the allocation of IP addresses and the network of domain name registrars who control all domain names but also delegate operation of the root servers. The root servers, the heart of the domain name system, are a collection of servers operated around the globe that manage all requests for information about the top-level domains (TLDs). TLDs are simply a means of organizing domain names into broad categories.

As of this writing, there are 14 generic TLDs in which entities or individuals can register secondary domains, in some cases subject to certain restrictions. They include the following:

.com for commercial sites,
.net for networks,
.org for nonprofit organizations,
.gov for U.S. federal government sites,
.edu for educational institutions,
.int for entities created by international treaties,
.mil for U.S. military sites,
.biz for businesses,
.info for general use,
.name for personal use by individuals,
.pro for professional fields such as lawyers and accountants (this TLD was still inactive as of February 2003),
.aero for the aerospace industry,
.coop for cooperatives, and
.museum for museums.

There are also more than 200 country code TLDs (ccTLD), based on the two-letter country codes for the worlds recognized nations. Examples include the following:

.us for the United States,
.uk for the United Kingdom,
.ca for Canada,
.mx for Mexico,
.de for Germany, and
.jp for Japan.

When you enter a domain name into your browser (for purposes of this example, I use www.example.com) here is—in theory—how the domain name system works to assure you get to the web site you want:

- Your browser communicates your request for www.example.com, via your Internet connection, to the Domain Name Servers designated for your use by your Internet service provider.
- Your service provider's domain name servers in turn ask the upstream DNS servers (and, if necessary, eventually, the root servers) to search their database for the IP address of the domain name servers that are authoritative for the TLD "com."
- Your query is then passed to the domain name servers for "com," which then search their database for the IP address of the domain name servers that are authoritative for the second-level domain "example" within the top level domain "com."
- Your query is then passed to the domain name servers for "example.com," which then searches their database for the IP address of the server that answers to the subdomain "www" within that second-level domain.
- Once it locates the correct IP address, it tells your Web browser what IP address to connect to, whereupon that server recognizes your request for a Web page and transmits the appropriate data back to your computer.

This is "in theory" because in reality, this process can be simpler, or more complex, depending on how your ISP chooses to manage its DNS requests. For example, some ISPs keeps a record of previous DNS requests in a "cache" file so that it can better manage time lag and server load issues by serving up IP addresses that it trusts are probably still correct because they were looked up a few hours earlier.

Domain Name Registration

The first challenge in registering a domain name is to identify a domain name that is suitable for your needs. Depending on the intended use of the domain name, there are many considerations, beginning with the choice of TLD that best suits your vision for your domain. Once you have decided on the TLD, you may have a choice of registrars delegated by ICANN to manage the process of domain name registration. For example, as of this writing there are several hundred ICANN-accredited registrars, not counting the designated registrars for all the country code TLDs.

Once you have selected a registrar, you will communicate to it what second-level domain you wish to register. In most cases, the registrar will check its records and determine whether the domain name requested is already registered or might have previously been reserved. Presumably you will have checked to see if there is a Web site already operating at the domain name you have selected; however, the absence of an active site is not determinative, because it is possible for a domain name to be registered but not in active use.

If the second-level domain name you desire has not been previously registered, you will likely be given the choice to register it. Upon providing contact and billing information, and paying the registrar's fee, of course, you will also be asked to provide the IP addresses of a primary and a secondary domain name server for your domain. Although some registrars offer the option of also hosting the domain on their own servers, you may need to have previously arranged with an ISP to establish the technical details necessary for operating DNS, Web, and e-mail services for your newly chosen domain name. If you have set up these services in advance, however, then it is possible to have your new domain fully functional within just a matter of minutes or hours of completing the registration process.

Internet Domain Disputes

Far and away the greatest amount of trademark-related controversy on the Internet concerns use of domain names. Because of both the value of trademarks themselves and the value of memorable domain names for maximizing the marketing and sales power of online operations, using popular trademarks as domain names has been an important issue for businesses beginning to make use of the Internet. Much to their consternation, however, many companies have attempted to register domain names related to their company name or their trademarks only to discover that someone else has already registered those domain names.

In the course of many legal disputes over domain names, some consensus among the courts has developed. Most courts have applied trademark law in much the same fashion as they would in any other trademark dispute. For example, marks are assessed for the extent to which they are fanciful, arbitrary, suggestive, descriptive, or generic. Disputes have also been judged on whether there is evidence of bad faith on the part of either party. In the Oppedahl & Larson case discussed earlier, there was no reasonable basis for the defendants to be making use of "Oppedahl" and "Larson" other than their desire to garner traffic attracted by someone else's mark.

The most common method of using a trademark in a domain name is the verbatim use of the mark in conjunction with the TLD, such as Pepsi.com. A related form of trademark infringement comes in dilution through the registration of similar domains, or domains containing misspellings or common typographical errors. Disputes over domains such as "amazom.com" (instead of amazon.com), gateway20000.com (instead of gateway2000.com), and micros0ft.com (microsoft.com with the second letter "o" replaced with a zero), have almost uniformly resulted in court decisions or settlements transferring domain ownership to the aggrieved party. These and other cases of infringement have resulted in a new area of law—and even of legislation—focused on resolving trademark-related domain name disputes.

Cybersquatting

In the mid-1990s, Dennis Toeppen registered some 250 domain names that were either similar or identical to popular trademarks, including deltaairlines.com, eddiebauer.com, neiman-marcus.com, northwestairlines.com, and yankeestadium.com. In two cases considered pivotal among domain name trademark disputes, *Intermatic Incorporated v. Toeppen* (1996) and *Panavision Int'l, L.P. v. Toeppen* (1996), the plaintiffs successfully forced Toeppen to relinquish control of the domains intermatic.com and panavision.com, respectively.

The Panavision case in particular illustrates how many of the cybersquatting disputes play out. In 1995, Toeppen registered the domain name www.panavision.com and created a Web site that contained photographs taken around the city of Pana, Illinois. When contacted by Panavision, a maker of motion picture cameras and photographic equipment, Toeppen offered to sell the domain name for $13,000. Panavision declined and brought suit under the Federal Trademark Dilution Act.

As discussed in an earlier section, the FTDA requires plaintiffs to demonstrate that their mark is "famous" and that the defendant is using a mark in commerce in a fashion that could cause dilution of the mark's distinctiveness. Although Toeppen claimed that his use of the mark was noncommercial, the court held that having offered the domain name for sale indicated that he intended that the domain name itself be a commercial offering.

In the Intermatic case, Toeppen originally operated a Web page at the intermatic.com address that described a piece of software he claimed to be developing called "Intermatic," later replacing it with information about Champaign–Urbana, Illinois, the community in which Toeppen lived. The Intermatic court held that despite these noncommercial uses, the registration of the domain name itself was dilutive of Intermatic's mark.

Anticybersquatting Consumer Protection Act

As the problem of cybersqatting grew throughout the 1990s, Congress responded in 1999 by enacting the Anticybersquatting Consumer Protection Act (ACPA), which amended the Lanham Act to include protections specific to Internet domain names. One change from past practice under the Federal Trademark Dilution Act, however, was the ACPA's removal of the requirement that the mark be used in commerce. This greatly expanded plaintiffs' ability to take control over domain names that had merely been registered but were not actually in use.

The ACPA states that cybersquatting occurs when the person registering a domain name containing a trademark "has a bad faith intent to profit from that mark" and "registers, traffics in, or uses" a domain name that is "identical or confusingly similar to or dilutive of that mark." The act

includes nine factors that courts may take into consideration when determining the existence of bad faith intent:

- The trademark or other intellectual property rights of the person, if any, in the domain name.
- The extent to which the domain name consists of the legal name of the person or a name that is otherwise commonly used to identify that person.
- The person's prior use, if any, of the domain name in connection with the bona fide offering of any goods or services.
- The person's bona fide noncommercial or fair use of the mark in a site accessible under the domain name.
- The person's intent to divert consumers from the mark owner's online location to a site accessible under the domain name that could harm the goodwill represented by the mark, either for commercial gain or with the intent to tarnish or disparage the mark, by creating a likelihood of confusion as to the source, sponsorship, affiliation, or endorsement of the site.
- The person's offer to transfer, sell, or otherwise assign the domain name to the mark owner or any third party for financial gain without having used, or having an intent to use, the domain name in the bona fide offering of any goods or services, or the person's prior conduct indicating a pattern of such conduct.
- The person's provision of material and misleading false contact information when applying for the registration of the domain name, the person's intentional failure to maintain accurate contact information, or the person's prior conduct indicating a pattern of such conduct.
- The person's registration or acquisition of multiple domain names that the person knows are identical or confusingly similar to marks of others that are distinctive at the time of registration of such domain names, or dilutive of famous marks of others that are famous at the time of registration of such domain names, without regard to the goods or services of the parties.
- The extent to which the mark incorporated in the person's domain name registration is or is not distinctive and famous.

The act then indicates that bad faith cannot be found if the defendant "believed and had reasonable grounds to believe that the use of the domain name was a fair use or otherwise lawful."

The recent dispute over the domain name Nissan.com gives some insight into how courts are applying the ACPA and other aspects of traditional trademark law analyses. The domain name is at the center of a dispute between Nissan Motor Co., Ltd., a popular car manufacturer, and Nissan Computer Corporation, a small business in North Carolina operated by Mr. Uzi Nissan. *Nissan Motor Co., Ltd v. Nissan Computer Corp.,* (2002). The court noted that the defendant registered the domain in 1994 and has been operating his firm under the Nissan name since 1985. Although the court found that the initial registration of the domain name was not in bad faith, the court did take issue with several aspects of Mr. Nissan's behavior, including his initial response to Nissan Motors complaint: an offer to sell the domain name for several million dollars. More recently, the court ordered Mr. Nissan to cease any commercial activities involving the site when it was discovered that Mr. Nissan was advertising automotive-related products on the site, despite his business being unrelated to automobiles (Nissan Computer Corporation Keeps Domain Name, 2002).

One of the innovative aspects of the ACPA is the way it deals with jurisdictional matters. Traditionally, legal disputes have always been subject to jurisdictional boundaries and national borders, and either the physical presence of the parties within those boundaries or evidence of the parties' contacts with the jurisdiction. Recognizing that in many cases cybersquatters go to some lengths to hide their identity or their location, the ACPA permits the trademark owner to take action against the domain name itself, rather than the domain owner personally. This permits aggrieved parties to locate the registrar and, although the registrar itself cannot be held liable for an infringing domain it permitted to be registered, the domain name can be attacked.

ICANN Domain Name Dispute Process

Although the ACPA's jurisdictional elements have simplified matters somewhat for trademark owners with domains that have been registered within the United States, the global nature of the Internet has required a less geocentric dispute resolution process. To that end, ICANN as developed a Uniform Domain Name Dispute Resolution Policy (UDRP; 2001), the provisions of which are in many respects similar to those of ACPA.

The UDRP process relies on third-party, private dispute resolution mechanisms rather than the more expensive prospect of litigation in a court of law. For trademark owners who are seeking a quicker and cheaper resolution of domain disputes, the UDRP route has proven to be extremely popular, even though no monetary damages or injunctions are available. Indeed, the only remedies available under the UDRP are the cancellation or transfer of the domain name. Thus, for instance where the injury to the trademark owner is more serious, traditional litigation may still be necessary.

The World Intellectual Property Organization (WIPO) has attempted to define cybersquatting and to establish guidelines for dispute resolution. In a 1999 report on the Management of Internet Names and Addresses, the WIPO delegates discuss at great length the phenomenon of registering trademarks as domain names with the intent of profiting from the ownership, either by capitalizing on traffic brought in through the confusion or by selling the domain to the trademark owner. In the end, WIPO's recommendations are similar to those contained in the ICANN UDRP and the ACPA.

The UDRP, WIPO's recommendations, and the operations (and even the very existence) of ICANN are currently the subject of tremendous international debate. Although the details of these arguments are too lengthy for inclusion in this brief chapter, many Web sites (including ICANNWatch.org and UDRPinfo.com) chronicle the legal, technical, and political arguments over this emerging field.

CONCLUSION

As the Internet becomes an even more critical channel for businesses to reach out to consumers, the value of a well-leveraged trademark has never been higher. At the same time, the pressures on trademark owners from infringing activities are requiring them to be ever more vigilant in their policing and prosecution of violators. In response to these pressures, courts and lawmakers have expanded and clarified traditional trademark protections, adding greatly to the remedies available to trademark owners who feel their rights have been violated. In this chapter, I have covered the fundamentals of trademark law and applied to the unique new situations presented by Internet technologies. Although the trademark space will continue to evolve, it is clear that the value of trademarks is as well-recognized as ever in the history of commerce.

If there is one conclusion to be drawn, it is that trademarks are a complex field of law and procedure, requiring expert guidance to provide maximum opportunity and protection. This chapter provides readers with a general overview of many current issues in trademark law, but it is not a substitute for qualified legal counsel. As I have noted repeatedly throughout the chapter, successful use of trademark law depends on many detailed analyses and procedural hurdles and requires a significant commitment of time and resources to take full advantage. Trademark law provides robust protections to those who, with assistance from talented counsel, seek to protect their goods and services in the marketplace.

GLOSSARY

Blurring At type of dilution in which the distinctiveness of a mark is weakened through its identification with dissimilar goods.

Deep linking Creating a Web page link that is tied directly to a document deep within the page hierarchy of a Web site, rather than simply linking to the main home page of the site.

Dilution A lessening of the value of a famous trademark caused by an unauthorized use of the mark, regardless of whether any actual confusion, deception, or mistake occurred.

Distinctiveness The ability of a mark to distinguish the goods and services of the mark own from the goods and services of another.

Domain name An alphanumeric electronic address on the Internet.

Famous trademark A court-determined trademark designation under 35 USC §1125(c).

Lanham Act Also known as the Trademark Act of 1946, it created a set of federal rules for governing the process of registering trademarks and established certain nationwide legal protections for trademark.

Likelihood of confusion The test of trademark infringement under the Lanham Act. A likelihood of confusion exists if a substantial number of reasonably prudent consumers are likely to be confused as to the source of the goods or services.

Infringement Use of a trademark in a way that is so similar to the owner's usage that an average consumer will be deceived, will mistake the infringing good for the original, or will experience confusion over the nature or origin of the product.

Initial interest confusion The use of another's trademark in a manner reasonably calculated to capture initial consumer attention, even though no actual sale is finally completed as a result of the confusion.

Intellectual property A set of legal theories that recognize property rights in intangible things such as ideas and intellectual creations.

Meta tags Hidden codes embedded in Web pages that contain key words related to the contents of a particular page, designed to be seen only by search engines.

Secondary meaning An association that has developed in the public's mind between the mark or trade dress of a product and owner of the mark or product.

Service mark A mark that is used to identify a service or the provider of a service rather than a tangible product, such as the name of a consulting firm or the name of a proprietary analytical process used by that consulting firm.

Tarnishment A type of dilution in which the mark is used in an unflattering light, such as by associating it with inferior or distasteful products or services.

Trademark A mark that is used by a manufacturer or merchant to identify the origin or ownership of goods and to distinguish them from others and the use of which is protected by law.

Trade dress The overall image of a product, composed of the nonfunctional elements of its design, packaging, or labeling, including specific colors or color combinations, a distinctive package shape, or specific symbols or design elements.

U.S. Patent and Trademark Office The federal agency charged with managing the nationwide issuance of patents and registration of trademarks.

CROSS REFERENCES

See *Copyright Law; Cyberlaw: The Major Areas, Development, and Provisions; Internet Literacy; Legal, Social and Ethical Issues; Patent Law.*

REFERENCES

Anticybersquatting Consumer Protection Act of 1999 (ACPA) (15 U.S.C. §1129). Retrieved May 9, 2003, from http://www4.law.cornell.edu/uscode/15/1129.html

Bicknell, C. (1999). Making a mint on wallstreet.com. *Wired News*. Retrieved December 1, 2002, from http://www.wired.com/news/business/0,1367,19285,00.html

Brookfield Communications, Inc., v. West Coast Entertainment Corp., 174. F.3d 1036 (9th Cir. 1999).

Coca-Cola Co. v. Gemini Rising, Inc., 346 F. Supp. 1183 (E.D.N.Y. 1972).

Hormel Foods Corp. v. Jim Henson Productions, 73 F.3d 497 (2d Cir. 1996).

ICANN Uniform Dispute Resolution Policy (2001). Retrieved December 1, 2002, from http://www.icann.org/udrp/udrp.htm

Intermatic Incorporated v. Toeppen, 947 F.Supp. 1227 (N.D. Ill. 1996). Retrieved December 1, 2002, from http://www.jmls.edu/cyber/cases/intermat.html

Nissan Motor Co., Ltd v. Nissan Computer Corp., 61 U.S.P.Q.2d 1839 (C.D. Cal., 2002).

Nissan Computer Corporation keeps domain name, for now (2002). *OfficialSpin.com*. Retrieved December 1, 2002, from http://www.officialspin.com/main.php3?action=recent&rid=405

Oppedahl & Larson v. Advanced Concepts, et al., Civ. No. 97-CV-1592 (D.C. Colo., 1997). Retrieved December 1, 2002, from from http://www.patents.com/ac

Panavision Int'l, L.P. v. Toeppen, 945 F. Supp. 1296 (C.D. Cal. 1996). Retrieved December 1, 2002, from http://www.jmls.edu/cyber/cases/panavis2.html

Playboy Enterprises, Inc. v. Terri Welles, 7 F. Supp. 2d 1098 (S.D. Ca. 1998), aff'd. in part, reversed in part, 162 F.3d 1169 (9th Cir. 2002).

Toys 'R' Us v. Akkaoui, 40 USPQ.2d 1836 (N.D. Cal. 1996).

Trademark Act of 1946 (also called the Lanham Act). (15 U.S.C. §1051). Retrieved December 1, 2002, from http://www4.law.cornell.edu/uscode/15/1051.html

United States Patent and Trademark Office (n.d.). *Trademarks*. Retrieved April 15, 2003, from http://www.uspto.gov/main/trademarks.htm

Zatarains, Inc. v. Oak Grove Smokehouse, Inc., 698 F.2d 786 (5th Cir. 1983).

Travel and Tourism

Daniel R. Fesenmaier, *University of Illinois at Urbana–Champaign*
Ulrike Gretzel, *University of Illinois at Urbana–Champaign*
Yeong-Hyeon Hwang, *University of Illinois at Urbana–Champaign*
Youcheng Wang, *University of Illinois at Urbana–Champaign*

INTRODUCTION

The travel industry is the world's largest industry, exceeding $4.5 trillion in gross output (World Travel and Tourism Council [WTTC], 2002). Recent reports from the World Travel and Tourism Council indicate that tourism employs over 198 million people worldwide, or approximately 7.8% of the global workforce. The emergence of travel as a significant economic activity began after World War II as travel became widely accessible to the general population. As shown in Table 1, very few people traveled internationally in 1950 as measured by today's standards. Yet, from 1950 to 1970, international travel exploded, increasing by more than 550%! This growth in international travel continued through the 1980s and 1990s to reach over 450 million visitor arrivals in 2001, representing over $260 billion (U.S. dollars) in expenditures. Since 1990 international travel has increased over 50%, and for a number of countries it has grown to be their largest commodity in international trade. Indeed, the travel industry now serves as one of the top three industries for almost every country worldwide (Goeldner & Ritchie, 2002).

Information technology has played a central role in the growth and development of the tourism industry. In the early years of mass global tourism (from the 1950s to the 1970s), the technology used was largely limited to computer systems that supported the internal functions of large operators in the transportation, hotel, and food service sectors. Also, a number of central reservation systems (CRSs) and global distribution systems (GDSs)—Sabre, Amadeus, Galileo and Worldspan—were developed to enable travel agencies (and other similar intermediaries) to directly access schedule and pricing information and to request reservations for clients. These intermediaries became the primary users of travel information systems, thus providing important links between travelers and industry players (World Tourism Organization Business Council [WTOBC], 1999).

During the late 1980s and early 1990s these systems and the information they contained emerged as important strategic tools, enabling system operators such as American Airlines and Hilton Hotels to grow and successfully position themselves within the overall travel market. The work of Mayros and Werner (1982) and Wiseman (1985) describes this significant development in the travel and tourism industry. An important characteristic of the growth of these systems was the inclusion of detailed behavioral information about each customer, including demographic characteristics, travel history, travel preferences, and responses to marketing/promotional programs. Armed with this information, existing systems were significantly enhanced and a variety of new systems were developed, which provided the basis for the emergence of a number of new approaches for managing tourism enterprises (Poon, 1993). As a consequence, the relationships of firms and organizations within the travel industry changed dramatically, placing emphasis on obtaining and distributing customer-related information as well as expanding strategic relationships in order to more fully exploit various business opportunities within the travel value chain.

The success of central reservation systems and global distribution systems paved the way for the Internet, enabling the travel and tourism industry to quickly exploit its many strengths. Indeed, in many ways the Internet is ideal for the travel and tourism industry. As a communication tool, it is simultaneously business- and consumer-oriented. The Internet is business-oriented because it enables businesses to communicate with potential visitors more easily and efficiently and allows them

Table 1 World Tourism Growth

Year	International Tourism Arrivals (millions)	International Tourist Receipts[a] (billions US$)
1950	25.3	2.1
1960	69.3	6.8
1970	165.8	17.9
1980	286.0	105.3
1985	327.2	118.1
1990	457.2	263.4
1995	550.3	406.2
1996	596.5	435.6
1997	618.2	439.6
1998	626.4	442.5
1999	650.4	455.4
2000	698.8	475.8

Source: World Tourism Organization (2001).
[a]International transport receipts excluded.

to better support and immediately respond to customer information needs through the provision of interactive travel brochures, virtual tours, and virtual travel communities. In addition, the Internet enables tourism-related enterprises to communicate with their partners more effectively in order to develop and design offers that meet the individual needs of potential visitors. The Internet is customer-oriented in that it empowers the user to easily access a wealth of information, enabling the traveler to almost "sample" the destination prior to an actual visit. Moreover, emerging ecommerce capabilities enable the traveler to make reservations, purchase tourism-related products, and share trip experiences with others.

Today, the Internet is one of the most important communication tools for travelers as well as travel and tourism enterprises. For example, recent studies by the Travel Industry Association of America (TIA) (2002a) indicate that almost one-third (31%) of American adults use the Internet to search for travel information and/or make reservations (see Table 2); for American travelers, this figure increases to 45%. Recent studies also show that the travel and tourism industry responded to the emergence of Internet-based technologies by adopting a number of new and innovative ways to communicate with consumers, as well as with other industry partners. As such, the travel and tourism industry is one of the most significant users of Internet technology as measured by the number of Web sites, Web pages, and online sales volume (Werthner & Klein, 1999). Indeed, recent studies of online spending show that consumers spent $19.4 billion in 2001 on U.S. travel sites, which accounts for approximately 36 percent of the $53 billion spent by consumers at all U.S. Internet retail sites (TIA, 2002b). Further, surveys of American convention and visitor bureaus show that essentially every bureau maintains a Web site, with some offering more advanced ecommerce capabilities, and surveys of the Internet indicate that over 75 million Web pages exist supporting the travel and tourism industry (Wang & Fesenmaier, 2002).

Table 2 Internet Use for Travel Planning: 1997–2002

Year	% of U.S. Adults	% of U.S. Travelers
1997	7	8
1998	18	21
1999	25	33
2000	30	40
2001	33	46
2002	31	45

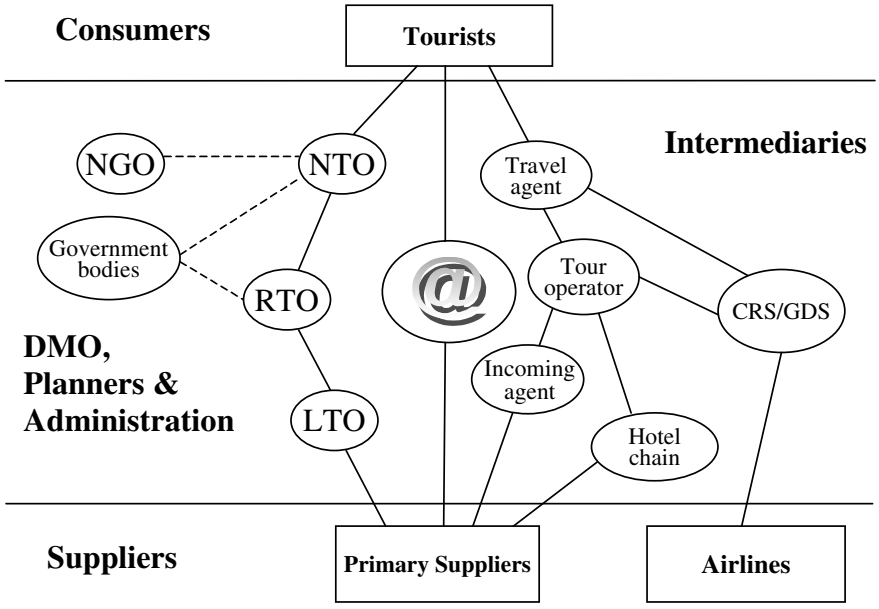
Source: TIA (2002a).

This chapter presents an overview of many of the uses of the Internet by travelers and the firms and organizations that compose the travel and tourism industry. The next section provides a brief synopsis of the structure of the industry and the role of Internet technologies, focusing on four major sectors: hotels, airlines, travel agents (and related intermediaries), and destination marketing organizations (DMOs). Following this introduction to the use of Internet-based technologies in travel and tourism, emerging marketing and management strategies are considered and various issues related to the development and use of Web sites and online management information systems are discussed. The subsequent section focuses on the role of the Internet from the travelers' perspective. A variety of consumer-related technologies are considered, including travel "brochure" Web sites, virtual tours, and mobile devices. As part of the discussion, changing patterns of use and their impact on the travel experience are considered. The last section of this chapter identifies five global technology-related trends affecting the future role of the Internet in travel and tourism. In addition, some possible ways in which the Internet will shape the future of the travel and tourism industry are discussed.

THE TRAVEL AND TOURISM INDUSTRY

Structure of the Industry and the Role of Information Technologies

The travel and tourism industry can be characterized as comprising all organizations that are involved in the production and distribution of travel and tourism products. It can be viewed as an umbrella industry (see Figure 1) containing a set of interrelated businesses, such as transportation companies, accommodation facilities, attractions, catering enterprises, tour operators, travel agents, and providers of recreation and leisure facilities (Werthner & Klein, 1999). To respond effectively to the dynamic character of the industry, information must be able to flow among the clients, intermediaries, and each of the suppliers involved in serving the clients' needs. As a result, information technology (IT) has become an almost universal distribution platform for the tourism industry. IT reduces the cost of each transaction by minimizing print, coordination, communication, and distribution costs. It also allows short-notice changes, supports one-to-one interaction with the customer, and enables organizations to reach a broad audience (Poon, 1993). However, the Internet has not affected all sectors equally. Certain sectors such as airlines have been aggressive adopters of technology, using it to help manage and streamline their



Note: NGO = Non-Governmental Organization NTO = National Tourism Organization
RTO = Regional Tourism Organization LTO = Local Tourism Organization
CRS = Central Reservation System GDS = Global Distribution System

Figure 1: The travel and tourism industry and the Internet. Source: Werthner & Klein (1999). Reprinted with permission.

operations and to gain strategic advantages (McGuffie, 1994). Others, such as the hotel sector, have been less enthusiastic and have only recently begun to take advantage of many of the benefits that the technology can bring (Connolly & Olsen, 1999). Many traditional travel agencies are also lagging behind other sectors in terms of technological adaptation, and it is increasingly evident that experienced consumers are often better informed than professional advisors. However, given the way in which information technology is reshaping the basic structure of both commerce and society in general, its importance to the success of all types of tourism companies can only grow in the future. As a result, tourism companies have changed dramatically the way in which they conduct their business and are under pressure to invest further in new technology in order to maintain their competitiveness.

Hotels

The hotel industry bases much of its distribution on direct contact with customers (WTOBC, 1999). Historically, hotels have distributed information through print-based media such as brochures, travel planners or regional guides, and received reservations by mail, phone, and fax. More recently, hotel rooms have been made accessible for booking through global distribution systems (GDSs) and through direct access to hotels using central reservation systems (CRS). However, such technologies have been inadequate as customers have traditionally not had access to these systems and travel intermediaries have experienced difficulty and delay in finding and booking appropriate hotels, whereas hotels have experienced high clerical costs attracting and processing bookings from customers. The emergence of new information and communication technologies (i.e., Internet technologies) presents new opportunities to make these processes more accessible and more efficient.

The use of the Internet in the hotel industry is growing exponentially and this enables hotels to reconsider the way they are doing business. Although the hotel sector overall has been slow to use the Internet as compared to other industry sectors (Connolly, Olson, & Moore, 1998), many hotel managers are becoming increasingly aware of the potential distribution, promotion, and interactive marketing advantages of the Internet. The Internet offers several advantages for hotels of all sizes. One of the advantages is increased effectiveness due to cost reduction and revenue growth. Another advantage is higher quality customer relationships due to the possibility of personal contact services and dialogue with the customer (Morrison, Taylor, Morrison, & Morrison, 1999; Sterne, 1999). For example, customers can answer questions about their personal preferences for rooms, and based on this information, a customer receives services at the hotel that are adapted to his/her preferences.

It is now generally agreed that Internet-related technologies are the single greatest force driving change in the hotel industry and will continue to have dramatic and sweeping implications on how hotels conduct business in the future. Hotels are expected to position themselves strongly on the Internet to take advantage of its distribution capabilities such as reach, content dissemination, feedback collection, interactivity, and one-to-one marketing. Further, current trends indicate that this greater involvement in IT by the hotel industry will increasingly encompass customer-centric approaches to capitalize on the cost structure and long-term potential of the Internet while at the same time differentiating products and building lasting value propositions.

Airlines

Air transportation systems worldwide are being dramatically affected by technological developments. Many of these developments focus on the use of Internet technology to improve the efficiency of operations (Sheldon, 1997). The first applications of computer technology to airline operations emerged in the 1950s when central reservation systems (CRSs) were designed. The primary function of computerized airline reservation systems was to simplify the process of booking flights by allowing travel agents to find relevant flight information and make reservations directly from their terminals without having to call airline reservations offices (Klein & Langenohl, 1994). Because of their many operational and cost-related advantages, CRSs became essential for the distribution of airline tickets through travel agencies.

Until the mid 1970s, airline computer reservation systems were used only for proprietary airline information and the major airline companies all had their own CRSs (Sheldon, 1997). Some of the CRSs were combined to become global distribution systems (GDSs), which provided multiple carrier information and constituted important electronic distribution channels. The major GDSs include Galileo, Amadeus, Sabre, and Worldspan, and these are now available through the Internet. These airline reservation systems provide a number of functionalities to travel agencies including flight schedules and availability, passenger information, fare quotes and rules, and ticketing. Most of the systems now also enable consumers to view schedules, fares, and fare rules and to book flights. In addition to developing reservation systems as the predominant distribution channel, many airlines have invested heavily in information systems to automate other areas of airline operations and management, which can be categorized into two sections: (1) systems for streamlining operations such as baggage and cargo handling systems, cabin automation, and safety systems, and (2) decision support systems to aid in decision-making related to flight scheduling and planning, crew scheduling and management, gate management, and departure control.

Travel Agencies/Online Intermediaries

Travel agencies are intermediaries that arrange and distribute travel information to individual travelers, with some agencies specializing in certain market segments or products. In addition, many travel agencies function as tour operators, designing their own package tours and selling them either directly to the traveler or through other agents. Travel agencies use information intensely and therefore need IT to process that information. In fact, information on travel products, destinations, schedules, fares, rates, and availability is their most important product and defines their existence. The more information a travel agency can access electronically, the more timely, accurate, and efficient services it can provide to its clients.

The most prevalent application of IT in travel agencies is the GDS terminal, which was first placed in travel agent offices by major airlines to facilitate airline bookings in the 1970s (Sheldon, 1997). GDS terminals are still the major information and booking tools used by travel agents for all types of travel products. However, the advent of the Internet has significantly changed the way travel and tourism products are distributed. Increasingly, consumers can access information online and travel agents have been forced to adapt to this change. Travel agents have an ambivalent relationship with the Internet because it can be a threat in that it makes products available directly to the consumer and yet it also provides additional business opportunities. Many travel agencies offer services on the Internet, giving them a much broader geographic consumer base than if they operated in traditional ways. They can receive bookings from clients through the Internet and can even book the passenger on flights without issuing paper tickets. Travel agents can also use the Internet as a research tool, and this might be particularly important in the future as some travel products become available only via the Internet. In addition, IT applications can be used by travel agencies to create value-added products or services through the online provision of their travel expertise in combination with the wider range of travel products and services available on the Internet. However, realizing that physical location is irrelevant in today's electronic marketplace, new types of travel agencies which exist only on the Internet, such as Expedia and Travelocity, are emerging and continue to raise the level of competition among travel agencies.

Destination Marketing Organizations

Destination marketing organizations (DMOs) are typically not-for-profit, small and medium-sized, information-intensive organizations that perform a wide range of activities to coordinate the diverse components of the tourism industry (Gartrell, 1988). DMOs act as a liaison, collecting and providing information to both the consumer and the industry in order to facilitate tourism promotion and development of a specific area.

In general, destination marketing organizations have been slow to adopt IT in their operations and marketing activities. It was not until the late 1980s that computer systems were adopted by the larger DMOs to enhance publications and information operations and, to a lesser extent, to support reservation services. During the late 1990s, as desktop computing technologies became more widely available, DMOs began to use IT more extensively. More and more DMO directors realized that Internet marketing was an inseparable, often critical part of their overall marketing endeavor. They have since developed a high level of interest in the Internet because the use of the Internet offers the potential to reach a large number of consumers at relatively low cost and provides information of greater depth and quality than traditional media. In other words, using Internet technology enables DMOs to promote their destinations' tourism products and services better, present associated organizations more equally, and collect customer information for effective customer relationship management (CRM). More importantly, the Internet allows them to improve business processes, conduct marketing research, provide customer service, and facilitate destination management and planning with less dependency on time and space.

Despite this potential, DMOs have not been particularly quick in establishing a sophisticated Internet presence. Several factors have contributed to the current status of DMOs' Internet strategy. One factor is related to the complex structure and relationships of the various constituents DMOs represent. The travel and tourism industry has a very complex structure, with a large percentage of the organizations classified as small businesses (Gartrell, 1988). Among the DMOs' constituents are marketing organizations at different levels, suppliers, and distributors. Although each entity maintains critical relationships with other entities in order to deliver the desired products and services, the enormousness of this diverse industry as well as the different benefits these constituents are seeking make the job of DMOs complex and difficult.

The unique characteristics of the Internet and the capabilities required of DMOs for implementing effective Internet marketing have also been identified as important influences. Marketing is a creative and adaptive discipline that requires constant regeneration and transformation in accordance with changes in the environment (Brownlie, Saren, Whittinton, & Wensley, 1994; Cronin, 1995). Thus, conventional marketing activities cannot be implemented on the Internet in their present form (Hoffman & Novak, 1996). For Internet marketing strategies to be successful, DMOs have to be aware of the capabilities and characteristics of the Internet and need to start developing new marketing concepts and paradigms, because the Internet presents a fundamentally different environment for marketing activities than traditional media (Connolly & Sigala, 2001). Despite these problems and challenges, DMOs have begun to recognize the opportunities that emerge from using the special features of the new medium, in particular the interactivity and multimedia capabilities it provides. As a result, the number of DMO Web sites is rising rapidly, offering online destination information with increasing quality and functionality.

Emerging Marketing and Management Strategies

The adoption of information technology has transformed the way in which the tourism industry conducts business. With the assistance of new technology, especially the Internet, new opportunities have emerged for tourism organizations, which enable them to market their travel products and services and manage their daily business activities more effectively. In particular, innovative marketing and management strategies have evolved in the areas of Web development, Web advertising/promotion, e-commerce activities, customer relationship management, and the use of online destination management systems.

Web Site Development

Hanson (2000) observed that there are three major stages in Web site development: (I) publishing, (II) database retrieval, and (III) personalized interaction. Stage I sites provide the same information to all users. Though a Stage I site can contain thousands of pages, pictures, sound, and video, it is limited in the dialogue it affords between the travel Web site and the user because it only broadcasts information from the Web site to the viewer. Modern Web tools make Stage I travel Web sites easy and inexpensive to develop in that almost any document can be converted and moved online.

With experience and investment, the travel organization moves to Stage II Web sites, which combine the publishing power of Stage I with the ability to retrieve information in response to user requests. The responses are dynamically turned into Web pages or e-mail. Interactivity and dialogue have started, although this activity is limited to a series of "ask–respond" interactions. However, the ability to use Web sites as points of access to images, sound, and databases is particularly valuable.

A Stage III travel Web site dynamically creates a page catering to an individual customer. It moves beyond an "ask–respond" interaction into a dialogue and may anticipate user choices and suggest possible alternatives. A Stage III travel Web site does more than just react to requests typed into forms or selected by clicking on an image. It requires the capabilities of Stages I and II plus the customization of content and functions to the needs of a specific user.

Destination marketing organizations use Web-based technologies in different ways and with varying intensity, owing to different backgrounds, financial resources, and marketing objectives (Yuan & Fesenmaier, 2000). Some DMOs are at a preliminary stage of utilizing Web technologies for marketing activities, and these Web sites are typically only used to broadcast information by providing brochure-like information. Others are more advanced and sophisticated in this regard, taking advantage of Web technologies to make business activities more effective and efficient, or even to re-engineer whole business practices. More advanced DMO Web sites typically include more sophisticated capabilities such as interactive queries and request forms, personalization, and recommendation functions.

Web Advertising/Promotion

The Internet is an almost pure manifestation of marketing principles and practices (Inkpen, 1998). It is a tourism marketer's dream because (1) it enables travel companies of different sizes to compete on more equal terms and (2) it allows a travel company to open up a direct and potentially personalized channel of communication with its customers. In other words, travel companies of all sizes are much more equal in their competition for consumers' attention on the Internet. Travel is the most important business on the Web in terms of the volume of advertising and promotion (eMarketer, 1999). It is also most likely to generate revenues and achieve profits through its Web presence. However, in order to be successful in Web advertising/promotion, tourism organizations have realized that the Web is a medium that combines the elements of other media. Hoffman and Novak (1996) describe the new form of communication that the Internet provides as an "interactive multimedia many-to-many communication model" where interactivity can also be *with* the medium in addition to *through* the medium. Travel and tourism fit especially well with interactive media because

they constitute an information intensive industry where transactions can be rather easily made online.

E-commerce Activities

Before the onset of the Internet, electronic commerce was usually conducted over a proprietary network connecting a group of organizations such as airline companies, travel agents, and hotels with each other through CRSs or GDSs. The nature of the transaction was purely business-to-business. Tourism businesses now use the Internet as a means of redefining their focus, creating new products, finding new distribution channels, and creating new markets. For example, the major airline sites now offer customer reservations, electronic tickets, seat selection, in-flight merchandise, and reward points; in addition, some of the airlines have enhanced their sites to offer lodging, transportation-package deals, and cruises through their alliance partners (Harrell Associates, 2002).

The use of the Internet in the travel and tourism industry has also been driven by the convergent forces of the shift of consumer behavior toward more intensive uses of online environments and the successful adaptation of marketing and sales strategies by the industry. For many consumers, online booking of travel is already the norm, and this can only be expected to strengthen in the immediate future. Travel is a product that online consumers want to purchase; indeed, according to Forrester Research (1999), it is the product that those who are online, but have not yet purchased online, want to purchase most. From the point of view of travel and tourism suppliers, however, there is some reticence from certain sectors of the industry, such as the cruise line industry, to compete directly with their traditional intermediaries by making the move to direct sales, whereas others, such as the airline industry, have embraced the new online channels with great enthusiasm.

E-commerce solutions are gaining momentum and are expanding beyond reservations to include supply-chain management (e.g., procurement), internal business applications through intranets, and other business-to-business transactions as well as business-to-customer sales. It is certain that the Internet will continue to become faster, more reliable and secure, and also more feature-rich. In addition, it will become more mobile through portable devices such as personal digital assistants (PDAs) and cell phones that can communicate with ambient intelligent devices embedded in appliances and will increasingly be enabled by speech, thus truly giving customers anytime, anywhere access in a format conducive to their needs.

Customer Relationship Management

Customer relationship management (CRM) is a managerial philosophy that calls for the reconfiguration of the travel organization's activities around the customer. Successful CRM strategies evolve out of the ability to effectively capture exhaustive data about existing and potential customers, to profile them accurately, to identify their individual needs and idiosyncratic expectations, and to generate actionable customer knowledge that can be distributed for *ad hoc* use at the point of contact (Newell, 2000). Further, the success of CRM initiatives is dependent on the ability to collect, store, and aggregate large amounts of customer information from various sources. One of the major driving forces of CRM using the Internet is the ability to target each individual interactively. With the Internet, individuals and travel marketers can interact, and this direct interaction creates customer value and sets the stage for relationship building. Travel marketers continue to seek ways for compiling accurate databases of personal information such as sociodemographic, socioeconomic, geographic, and behavioral characteristics for potential customers. Such a database creates a wealth of relationship marketing opportunities. Crucial to the establishment of such comprehensive customer databases is the ability to use software agents, without human intervention, to collect, categorize, and store large amounts of personal customer information in a cost effective manner for later data mining. A second important issue is the ability to collect the desired information directly from the primary source rather than having to purchase it from secondary sources such as travel and tourism consultants.

Online Destination Management Systems

The term "destination management system" (DMS) has come into use in recent years to describe the IT infrastructure of a destination marketing organization and may be defined in a number of ways depending on the capabilities of the system. Increasingly, a DMS is regarded as having to support multiple functions. An integrated DMS supports not only the DMO's Web site, but also a wide range of other promotion, marketing, and sales applications (Sheldon, 1997). These might include the design and production of printed materials, tourist information center services (for information and reservations), call center services, kiosks, database marketing, project/event management, and marketing research. DMSs can greatly enhance a travel destination's Web presence by integrating information from various suppliers and intermediaries and are increasingly used as the informational and structural basis for regional Web portal sites.

TRAVELERS AND THE INTERNET

Internet technologies have not only changed the structure of tourism and its related industries. They have also had a profound impact on the way consumers search for tourism information, construct and share tourism experiences, and purchase tourism products and services. In contrast to many consumer goods and services, the consumption of tourism experiences involves often extensive pre- and post-consumption stages, in addition to the actual trip, which itself can spread over several weeks (Jeng & Fesenmaier, 2002; Moutinho, 1987). These stages of the tourism consumption process are typically information-intensive, and Internet-based technologies have come to play a significant role in supporting consumers throughout this multistage process. The specific ways in which the various technologies are used in the different stages depend on the particular communication and information needs they are expected to serve (see Figure 2). For instance, Internet technologies are used in the preconsumption phase to obtain information necessary for planning

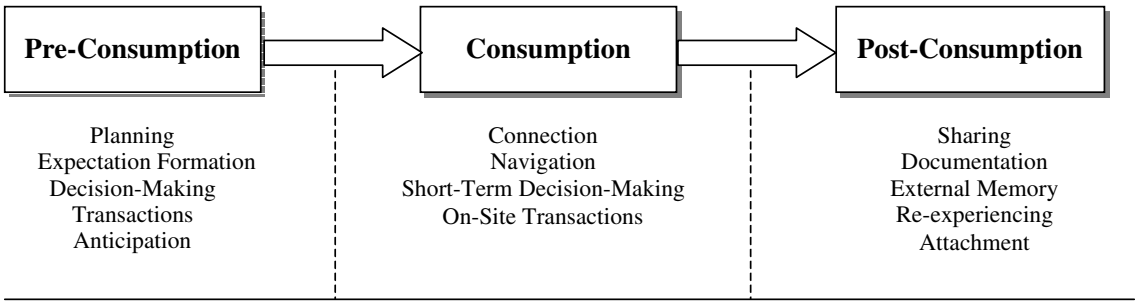


Figure 2: Communication and information needs in the three stages of tourism consumption.

trips, formulate correct expectations, and evaluate, compare, and select alternatives, as well as to communicate with the providers of tourism products and services to prepare or execute transactions. In contrast, the functions served by technologies during the actual consumption of tourism experiences are more related to being connected and to obtaining detailed information relevant at a specific place and moment in time. During the postconsumption phase, Internet technologies are used in ways that allow sharing, documenting, storing, and reliving tourism experiences, as well as establishing close relationships with places, attractions, or product/service providers, as in the case of Frequent Flyer programs. For example, e-mail will typically be used in all stages, but mainly to obtain information or make reservations in the preconsumption phase, to stay connected with family and friends while traveling, and to share pictures and stories with members of one's travel party or other individuals after concluding a trip. Although all Internet technologies are probably used by travelers at some point or in some way, several applications have been identified as being of particular importance in the context of tourism experiences. The following provides an overview of these technologies, how they tend to be used, and the impact they have on consumers during the various stages of the tourism consumption process.

Preconsumption

It is in the initial phase of the tourism consumption process that most of the impacts related to Internet-based technologies are currently experienced. Consumers use the Internet and its diverse applications in this first stage of the tourism experience to gather information, formulate expectations, inform/support their decision-making, and reserve or purchase the various components (transportation, accommodation, etc.) to be consumed during their trips.

Brochureware

Brochureware refers to Web sites or Web pages created by transferring the contents of printed tourism brochures directly to digital environments. Brochureware was one of the first Internet applications made available to tourism consumers, owing to the fact that tourism businesses quickly recognized the value of the Internet as a powerful publishing medium. Web sites designed as brochureware represent the simplest form of Web design and completely ignore the content presentation and communication possibilities the World Wide Web offers (Hanson, 2000). Nevertheless, brochureware is the most common way in which tourism information is currently made available to consumers and, consequently, constitutes an integral part of tourism-related online experiences. Despite their obvious limitations, digitized tourism brochures on the Internet still support consumers in that they enable potential travelers to browse and evaluate tourism products without temporal or spatial limitations. Furthermore, even the very basic implementations of brochureware make use of hypertext and provide hyperlinks that allow consumers to move through online tourism information in ways that are typically not supported by printed brochures. Brochureware is expected to give way to more interactive forms of Web site content presentation as more and more tourism businesses recognize the value of engaging consumers in experiential ways.

Virtual Tours

Virtual tours are tools that enable the potential consumers of tourism products to explore and immerse themselves within an interactive Web environment in order to gain the needed experiential information about a destination or tourism establishment (Cho & Fesenmaier, 2001). The term "virtual tour" is widely used on the Web and can range from a series of pictures or slide shows to streaming video and highly interactive virtual reality settings. The realism provided through virtual tours creates immersion, which, in turn, leads to immediate, direct, and real experiences that generate a strong sense of presence. As a result of this telepresence experienced through virtual tours, consumers are able to construct a more vivid picture of the tourism product and are therefore more likely to reach well-informed decisions. Thus, the significance of virtual tours in the context of tourism lies in providing consumers with an opportunity for "product trial" before the actual purchase. Tourism products are, in large part, experience-oriented intangible goods (Vogt & Fesenmaier, 1998) that are typically consumed at a place far away from the point of purchase and often cannot be experienced without being consumed in their entirety. Consequently, product trial is usually not available to the potential consumers of tourism products. However, tourism bears many risks because its components are consumed in unfamiliar environments, constitute a significant expenditure for most consumers, and typically entail high involvement from the part of the consumer. Given

the limited opportunity for prepurchase trial in the context of tourism, virtual tours, with their ability to represent tourism products and services in more realistic and dynamic ways than other promotional materials, play a crucial role in offering rich travel information.

Travel Decision Support Systems

The term travel decision support systems (TDSS) refers to information systems designed to simplify the travel-decision-making process and support consumers in the various steps involved in planning trips. Trip planning is a very complex process that consists of a number of decisions, which often condition each other (Dellaert, Ettema, & Lindh, 1998; Jeng & Fesenmaier, 2002; Woodside & MacDonald, 1994). Also, each decision requires different kinds of information; thus, separate information search activities are necessary. This increases the cost of the information search process for the consumer and can often lead to information overload (Good et al., 1999; Hibbard, 1997). The main function of a TDSS is to identify and present certain tourism products and services in accordance with consumer preferences, thus reducing the number of alternatives that the consumer would have to evaluate. The functionality of such systems ranges from more sophisticated search engines and intelligent-agent-supported information retrieval to true recommendation systems that enable the consumer to identify destinations or tourism products of interest (Vanhof & Molderez, 1994). The use of a TDSS generally requires specifying preferences such as desired date of travel and preferred activities. In their most advanced form, these systems try to mimic face-to-face human interactions that could occur between a consumer and a travel agent. In such cases, the TDSS is typically referred to as a travel counseling system (Hruschka & Mazanec, 1990). Current developments in the TDSS area focus on increasing the ability of such systems to capture consumer preferences and adapt information accordingly, as well as to learn from past interactions and support group decision-making (Delgado & Davidson, 2002; Hwang & Fesenmaier, 2001; Loban, 1997; Mitsche, 2001; Ricci & Werthner, 2001). Consumers can take advantage of such systems and benefit from the customized information presented to them at various stages in the trip planning process. However, the currently available TDSS versions are still more capable of supporting consumers with a clear understanding of what they desire than helping individuals with only a vague idea of what they are looking for. It is expected that these systems will play an increasingly important role in travel planning as they become more human-centric in design and truly adaptive with respect to the needs of consumers.

E-commerce Applications

E-commerce applications are technologies that support online transactions between the providers and consumers of tourism products and services. Online reservation and payment options were quickly adopted by many suppliers and consumers and led to the emergence of tourism as one of the most important e-commerce categories. It can be argued that the reasons for this rapid adoption of e-commerce in tourism lie in the particular fit between the characteristics of tourism products and the capabilities of e-commerce applications. The purchase of tourism-related products and services typically involves the movement of information rather than the physical delivery of goods and is often concluded through credit card payments. Also, the complex and strictly hierarchical tourism distribution system of pre-Internet times led to enormous information asymmetries and little choice for consumers in terms of where or how to acquire tourism products. With the introduction of e-commerce, consumers were provided with a new means of buying that they were eager to adopt because it not only offered more choices through direct links to the many geographically dispersed industry players or new intermediaries such as online travel agencies, but also catered to consumers by providing new levels of convenience. E-tickets, for example, are a direct result of e-commerce initiatives and have tremendously simplified travel, especially business travel. Current developments in the area of e-commerce focus primarily on raising security and privacy standards to ensure safe and smooth transactions. In addition, efforts are being undertaken to make a wider range of tourism products available online. Destination management systems are becoming more widespread and promise more extensive e-commerce adoption, thus providing consumers with greater access to an increasing variety of tourism products and services.

Online Customer Support

Online customer support is a summary term for Internet technologies that allow consumers to contact the suppliers or distributors of tourism products if additional information or other forms of assistance are needed. Applications that provide consumers with the means to communicate faster and more easily in order to get support are especially important in the context of tourism. First, the special nature of tourism products and services (see Kotler, Bowen, & Makens, 1999, for a detailed description) makes them more likely to require additional support. It is very difficult to describe the many experiential aspects of tourism accurately; thus consumers often require further interpretation or more detailed explanations. Also, tourism products and services are usually purchased or at least reserved long before they are consumed, and many things can potentially happen during this extensive period of commitment between travel- and tourism-related businesses and consumers.

Second, geographical and cultural distances between travel and tourism suppliers and consumers render communication through traditional means very difficult. The Internet, on the other hand, provides consumers with fast, easy, and cost-effective ways of contacting the providers of travel and tourism-related goods and services. Technologies of support include "Frequently Asked Questions" sections on Web sites, online request forms, bulletin boards, Internet phones, e-mail, real-time chat options, and instant messaging applications. These technologies are currently used in very passive ways, which means consumers are required to initiate the contact. However, a growing number of tourism organizations is adopting more proactive approaches to online customer support. They provide consumers with active assistance either through automatic e-mail up-dates or by monitoring the behavior

of Web site visitors and offering real-time assistance through instant messaging or chat if they recognize search or click patterns that are typically associated with a need for help, such as seemingly uncoordinated click-streams. This innovative use of online customer support technologies actively encourages consumers to communicate with customer representatives and has the potential to prevent confusion or misunderstandings instead of following the traditional model of solving problems after they have occurred.

During Consumption

Internet technologies are used during the actual trip mainly for travelers to stay connected and to obtain *en route* information if, and only if, the need arises. The spread of Internet cafés at tourist destinations, the growing number of accommodation establishments offering (often high-speed) Internet connections, and the recent efforts of airlines to provide in-flight Internet access to travelers indicate that a substantial need for these kinds of information and communication links exists. *En route* Internet access means anywhere-and-anytime availability of tourism-related information for consumers. Therefore, many of the trip planning and information gathering tasks of travelers could shift from preconsumption to during consumption and make travel much more spontaneous if the Internet becomes more widely available to the traveling public.

Mobile technologies play an increasingly important role in tourism due to their ability to provide travelers with wireless and, thus, instantaneous and pervasive Internet access. Handheld devices such as personal digital assistants (PDAs) and cellular phones supported through a wireless application protocol (WAP), a global system for mobile communication (GSM), and short message service (SMS) allow travelers to take full advantage of the Internet while on the road. More ambitious developments of mobile technology go beyond simple access by providing real-time location-based services (Eriksson, 2002; Oertel, Steinmüller, & Kuom, 2002). Empowered by geographical information systems (GIS) and global positioning system technology (GPS) in combination with information available on the Internet, these advanced mobile applications identify the traveler's location in space and the spatial context of this position. This information is then used to generate personalized assistance in the form of location-specific and time-sensitive information. Many advancements related to mobile technologies are spurred by needs that directly arise from information and communication problems encountered during travel. Projects such as CRUMPET—creation of user-friendly mobile services personalized for tourism (Poslad et al., 2001; Schmidt-Belz, Makelainen, Nick, & Poslad, 2002)—and the development of wireless-based tourism infrastructures, for instance the ambient intelligence landscape described by the Information Society Technologies Advisory Group [ISTAG] (2002), are two examples of the many efforts undertaken at the juncture of mobile technology and tourism; yet they represent developments with important implications for the future of the entire Internet.

Postconsumption

The postconsumption stage in the context of tourism, involves treasuring souvenirs, remembering special moments, reliving an experience through photographs, sharing travel stories, and often developing a strong sense of attachment to a specific destination. Internet technologies play a significant role in these post-trip activities and have started to significantly influence memory practices as they relate to tourism.

Virtual communities are an example of Internet applications that provide consumers with support during the postconsumption phase. The term *virtual community* describes a group of people who are connected through computer-mediated communication technologies and share interests and feelings in cyberspace (Rheingold, 1994). Virtual travel communities, then, are communities facilitated by computer-mediated communication that allow members to conduct various types of travel-related tasks, such as obtaining travel information, maintaining connections, finding travel companions, or simply having fun by telling each other interesting travel experiences and stories (Wang, Yu, & Fesenmaier, 2002). Consumers can use these virtual travel communities to post photographs and stories/testimonials of their trip(s) on the community Web site, where they serve as information to other consumers. In addition to this purely functional aspect, virtual travel communities offer opportunities for members to fulfill hedonic, psychological, and social needs. Sense of belonging, fun, and self-identification are only a few of the benefits that can be derived from online community membership. In the context of tourism, the most important function virtual travel communities serve is the extension of travel/tourism-related experiences beyond the actual trip. Used as digital substitutes for traditional photo albums, the digital images uploaded onto community Web pages and discussion boards help recall aspects of trips and assist consumers in constructing memories of vacations. The travel stories and discussions that can be found in such communities mimic real-world storytelling activities typical of this last stage of the tourism consumption process.

Tourism experiences are an integral part of the collective memory of families and peer groups and, thus, require sharing. In contrast to traditional conversations about the adventures, fun events, or other types of memorable moments of past trips, communication about travel experiences in virtual communities takes place with an audience that has a very tailored interest in the topic and often resides outside the boundaries of one's usual social circle. Also, the information posted by consumers in the course of the postconsumption recollection of the travel experience serves as valid information for consumers in earlier stages, thus closing the loop of the tourism information cycle. The information contained in virtual travel communities is especially valuable, as it represents personal accounts of probably alike consumers with actual product knowledge and no commercial interests. Thus, virtual travel communities can serve as a vehicle to control the quality of travel products and services through consumers' evaluation and ratings of a wide range of travel products and services. However, increasing numbers

of advertisers and tourism businesses are discovering virtual travel communities as particularly suitable vehicles to communicate messages to and establish relationships with specific target markets. Thus, it is unclear if virtual travel communities will remain in the control of consumers and will continue to be used as personal tools to share tourism-related information or if their focus will shift toward more commercial content and usage as promotional tools.

Impacts of Internet Technology on Travel Behavior

The Internet has had and will continue to have a tremendous impact on the way consumers search for, purchase, consume, and remember tourism experiences. However, the Internet is not the only channel through which consumers obtain information, communicate, or complete transactions. Rather, it is one of many options currently used by tourism consumers. Word of mouth, for instance, remains the most popular way of gaining access to first-hand knowledge about travel destinations and tourism experiences. Also, travel magazines and movies continue to be significant sources of inspiration. It seems that the concept of the "hybrid" consumer who uses many media and technologies simultaneously (Wind, Mahajan, & Gunther, 2002) is especially applicable to tourism. Thus, the Internet has not replaced traditional channels but has placed additional options in the hands of consumers. Nevertheless, many current technology developments aim at convergence and the creation of one channel that can satisfy all information search, transaction, and communication needs and, therefore, this situation might change in the near future.

In general, the Internet and its many different applications have provided consumers with an incredible number of choices, opportunities for comparison shopping and much more control over many processes related to the consumption of tourism experiences. The success of auction models such as Priceline.com, where consumers name prices rather than accepting the industry-imposed price, indicates that the market has shifted from a supplier- to a consumer-dominated market.

Tourism has always been characterized by many alternatives. Yet many consumers lacked the necessary information to take advantage of the variety of tourism offers available. The Internet has, to a large extent, closed this information gap. Internet consumers are much more informed, and this new level of information and knowledge among "new" consumers has opened up many choices. The larger extent and different, more experiential nature of information available to consumers has also led to more accurate expectation formation and more informed decision-making, both of which are especially important for the consumers of information-intensive, intangible, and high-involvement tourism products and services. On one hand, the ease with which information can be made available online has placed this abundance of information at the disposal of the consumers. On the other hand, it causes situations of severe information overload and leads to many concerns about trust. The Internet facilitates information representation and distribution; however, it also makes it more difficult for consumers to identify false information. Many tourism businesses are very small and operate thousands of miles away from where their customers live. It is, thus, extremely difficult for consumers to verify their existence, not to speak of the nature and quality of their business practices. Consequently, the consumers of travel- and tourism-related products can be expected to continue relying on offline and online intermediaries as well as official Web portals such as destination marketing sites to obtain reliable and trustworthy information about tourism establishments.

The Internet has increased the speed with which information moves between tourism suppliers and consumers, and many consumers have started to expect instantaneous information and support with respect to all aspects of their trips. The Internet has influenced not only perceptions of speed but also the extent of personalization expected by the consumers of tourism information and products. These new expectations in terms of speed and personalization spurred by the Internet represent an enormous challenge for the tourism industry and leave many consumers disappointed with the level of service they receive.

One of the main advantages of the Internet is that it provides consumers with the opportunity for anytime-and-anywhere access to information. Many aspects of a trip that had to be planned well in advance can now be finalized while on the road. Internet technologies that provide consumers with such *en route* access to information have the potential to significantly transform trip planning and influence travel patterns. Recent trends indicate that travel is in the process of becoming more spontaneous and that many travelers choose to travel to destinations that they would not traditionally have considered because of the high risk. One can only speculate about the impact of this trend on tourism in the Internet era as extensive planning is an integral part of the tourism consumption process and careful preparation is often essential to the success of a trip (and sometimes even crucial to the survival of the travel party).

The advent of the Internet has brought about many dystopian fears related to the future of tourism. Many predicted the end of travel *per se* and pictured tourism experiences as being confined to virtual reality simulations. The future of tourism on the Internet from a utopian perspective of course looks much brighter. Such a perspective interprets Internet experiences as a substitute for travel and, thus, a great opportunity for individuals with disabilities or other constraints that limit them from traveling. Neither scenario has been realized so far. Internet technologies have discouraged some types of travel and encouraged others. The experiences presented on the Internet are far from constituting real substitutes for actual travel, but they provide invaluable information about accessibility and allow individuals with special needs to prepare more accurately for real-world trips. Consequently, when analyzing the influence of Internet technologies on consumer behavior in the context of tourism, one has to constantly remind oneself that the Internet is still in the process of becoming and that its impact on the consumers of travel and tourism products and services can be partly grasped but not yet fully understood.

TRAVEL AND TOURISM FUTURES

For the tourism industry, the Internet is clearly the biggest opportunity but simultaneously the biggest challenge. Looking to the future, of course, poses many questions; however, there are a number of general trends pointing in the direction of the future development of travel and tourism. In this section five technology-related trends will be considered: (1) the continuing speed and sophistication of information technology; (2) the continuing growth in the use and uses of information technology in tourism; (3) the changing forms of information technology as a medium for communication; (4) the emergence of a new tourism consumer; and (5) the emergence of experience as the foundation for defining tourism products. The following will identify and briefly discuss each trend, focusing attention on its impact on the tourism industry.

Trend #1. The Continuing Speed and Sophistication of Information Technology

The personal computer has only recently celebrated its 20th birthday. In a recent article in *PC Magazine,* the personal computer (PC) was described as one of the most profound inventions in the history of mankind (Miller, 2002). From its inception in 1981 the development of computer technology has been shown to follow Moore's Law—that chip density and therefore the speed of computers will double every 18 months. Thus, computers and computer technology have grown from the very "primitive" Radio Shack TRS-80 and the IBM 4.7 MHz 64K RAM 8088 processor, an operating system called DOS, and software called VisiCalc and Wordstar to today's 2.5-GHz machines with over 1 GB of RAM. Over the years, various competitors have infused the market with a variety of innovations focused on expanding the power of the machine and the ability of the system to address workplace needs and encouraging society to think/dream about what might be in the future (Miller, 2002). As the systems grew more sophisticated and powerful and arguably more human-centric, the power of the network was recognized and spurred even greater innovation. In 1990 the World Wide Web was born, along with a new generation of innovators seeking to build an information infrastructure that could enable individuals to collaborate from distant locations. The outcome was MOSAIC and a decade of unparalleled innovation and "build out" in information infrastructure. This new orientation also led to the development of a variety of computer-enabled devices, such as cell phones and personal digital assistants, which are now beginning to pervade human society (Norman, 1999).

A number of scholars have recently reflected on the progress of computer technology and have concluded that there is much to be accomplished before computers/information technologies can truly enable society to benefit from their power. In *The Unfinished Revolution,* Dertouzos (2001) argued that "the real utility of computers, and the true value of the Information Revolution, still lie ahead" (p. 6). He suggested that over the past 20 years society has evolved to "fit" around computers and that the productivity gains from computer technology have been "more hype than reality." Supporting this argument, Norman (1999) suggested that the real benefits will be realized only when computer technology becomes more human-centered, that is, when technology adapts to the needs and lifestyles of human beings. They argue that information appliances—computer systems that focus on specific tasks and are connected through the Internet or wireless technology—are the basis of human-centric and thus "invisible" computing. It appears that the focus of emerging technology is on empowering the individual within the framework of the human experience rather than defining human behavior around the needs of computer designers.

Examples of emerging technology that is beginning to be used in the travel and tourism industry include travel recommendation systems, virtual reality, and travel guidance systems. A number of basic recommendation systems are available in a variety of travel-related Web sites including Ski-europe.com, Travelocity.com, and TIScover.com. Indeed, virtual reality and related technologies are evolving sufficiently to enable travelers to sample/experience the destinations prior to the actual trip. In addition, GPS-supported travel guidance systems, which once were considered exotic, are actively marketed for automobiles.

Trend #2. Continuing Growth in the Use and Uses of Information Technology in Tourism

The number of Internet users continues to grow worldwide and as a result, the Internet's potential as a marketing medium has expanded greatly and continues to expand. Internet revenues show a robust growth from about $7.8 billion in 1997 to a projected $34 billion in 2002 (Forrester Research, 1999). The characteristics of the Internet have considerably changed over the past five years. The present Internet users have become older, considerably less male, and relatively less educated and are more likely to have families, more likely to live in geographic regions corresponding to the general U.S. population distribution, and more likely to represent a broader range of occupational categories than their predecessors.

The three leading uses of the Internet cited by all users were information gathering, searching, and browsing. Whereas male users are more likely to use the Internet for information gathering, work, and shopping, female users are more likely to use the Internet for searching, browsing, and education (Pitkow, 1997). According to the latest FIND/SVP study (1994–1997), the Internet is considered indispensable by 73% of users to describe the impact of the Internet on their lives. An overwhelming majority (over 87.9%) uses the Internet for personal reasons such as e-mail and instant messaging; 76.3% search for news, product information, educational information, or entertainment (University of California, Los Angeles [UCLA], 2001). Business use was reported by 60% of all adult Internet users (21.7 million users) and includes applications such as file transfer, searching for business news, conducting business research, and shopping for business-related goods and services. Further, these studies indicate a substantial increase in online use in terms of the average

online session length, the number of hours spent online per week, and the number of Web sites visited regularly, registered with, and paid for by users (FIND/SVP, 1994–1997; UCLA, 2001).

The Internet and travel industry partnership has proved beneficial to both industries (Gretzel, Yuan, & Fesenmaier, 2000; Hardie, Bluestein, McKnight, & Davis, 1997; Jupiter Communications, 1997). Travelers' usage of the Internet has grown progressively from 1996 to 2001. The incidence of Internet use among American travelers has increased from 28 million Internet users in 1996 to 110 million in 2001 (TIA, 1998; 2002a). The Travel Industry Association of America Report on Technology and Travel for the year 2000 (TIA, 2000) reported that 89% of Internet users took at least one trip (for business or leisure) of 100 miles or more, one way, away from home during the year and 44% of Internet users were frequent travelers (who had taken five or more trips in the previous year). According to the 2001 National Travel Survey, 68% of current Internet users used the Internet to make travel plans (TIA, 2002a). Travel plans include activities such as getting information on destinations and checking prices and schedules. This number was up from 27% of Internet users in 1997 and 10% in 1996. Among Americans who did travel planning over the Internet in 1997, 7% did all of their travel planning over the Internet, 16% did most of their travel planning over the Internet, and nearly one-third used the Internet half the time for collecting travel information. The TIA report also indicated that in 2001, one third (33%) of American travelers who are online indicated they actually booked or made reservations online. The large majority of these travelers purchased airline tickets (80%), reserved a hotel room (62%), or rented a car (46%). In addition, many online travelers purchased tickets for cultural events (27%) and/or aamusement parks (14%). As shown in Table 3, these numbers were slightly lower in 2002, possibly reflecting the uncertainty caused by the economic downturn and political instability through terrorism.

Table 3 Travel Products/Service Purchased Online In 2002 (among 39.0 Million U.S. Travelers Who Have Internet Access and Who Booked Travel Online)

Travel Products/Services	2001	2002
Airline ticket	80%	77%
Overnight lodging accommodations	62	57
Rental car	46	37
Tickets—cultural event	27	25
Travel package	13	21
Tickets—spectator sporting event	16	18
Reservations for personal sports (like golf/skiing/water sports)	NA	13
Tickets—amusement park	14	12
Tickets—museum or festival	11	11
Cruise	NA	6
Tickets—tour or excursion	NA	6

Source: TIA (2002a).

Trend #3. Changing Forms of Information Technology as a Medium for Communication

Industry experts have increasingly questioned whether the Internet is different from other media and if it needs to be addressed in new ways using new strategies (Godin, 1999; Hoffman & Novak, 1996; Zeff & Aronson, 1999). The Internet is special in that both consumers and firms can interact with the medium, provide content to the medium, communicate one to one or one to many, and have more direct control over the way they communicate than using other media. When everyone can communicate with everyone else not only the old communication models become obsolete but also the communication channels that are based on them (Evans & Wurster, 1999).

In contrast to traditional media, the Internet combines and integrates the following functional properties: (1) information representation; (2) collaboration; (3) communication; (4) interactivity; and (5) transactions. As a consequence, Internet communication can be much more holistic than communication through traditional media. The Internet can simultaneously integrate informational, educational, entertainment, and sales aspects; this flexibility makes the Internet rich and appealing but also very complex and difficult to deal with. Interactive media such as the Internet call for interactive marketing. "The essence of interactive marketing is the use of information *from* the customer rather than *about* the customer" (Day, 1998, p. 47). It differs from traditional marketing because it is based on a dialogue instead of a one-way communication, and it deals with individual consumers instead of mass markets (Parsons, Zeisser, & Waitman, 1998). Using these capabilities of the Internet may lead to deeper relationships with customers and greater personalization of goods and services. Travel and tourism fit especially well with this interactivity aspect of the Internet because they are an information-intensive and experience-based industry.

The Internet enables destination-marketing organizations to blend together publishing, real-time communication, broadcast, and narrowcast (Hoffman, Novak, & Chatterjee, 1995). It is a medium that attracts attention and creates a sense of community. It is a personal medium, an interactive medium, and a niche and a mass medium at the same time (Schwartz, 1998). In contrast to traditional media, the trade-off between richness and reach is not applicable to the Web. Evans and Wurster (1999) define richness as the quality of the information presented (accuracy, bandwidth, currency, customization, interactivity, relevance, security). Reach refers to the number of people who participate in the sharing of that information. The trade-off between richness and reach leads to asymmetries of information. Thus, when destination-marketing organizations are able to distribute and exchange rich information without constraint, "the channel choices for marketers, the inefficiencies of consumer search, the hierarchical structure of supply chains, the organizational pyramid, asymmetries of information, and the boundaries of the corporation itself will all be thrown into question" (Evans & Wurster, 1999, p. 37).

Trend #4. Emergence of a New Tourism Consumer

The Internet changes how people communicate and exchange information. The resulting abundance of information and ease of communication have led to profound changes in consumer attitudes and behavior. What makes new consumers "new" is that they are empowered by the Internet, which provides them with easy and cheap access to various information sources and extended communities (Windham & Orton, 2000). New tourism consumers are well informed, are used to having many choices, expect speed, and use technologies to overcome the physical constraints of bodies and borders (Poon, 1993). Lewis and Bridger (2000) describe the new consumer as being (1) individualistic; (2) involved; (3) independent; and (4) informed. The Internet is a highly personalized medium and new consumers expect marketers to address and cater to their complex personal preferences. Consequently, new tourism consumers are "in control" and have become important players in the process of creating and shaping brands.

New tourism consumers are also very independent in making consumption decisions but, at the same time, like to share stories about their travel experiences with members of different communities. Stories can convey emotional aspects of experiences and product/service qualities that are generally hard to express in writing and, consequently, are rarely included in traditional product descriptions. Storytelling is an important means of creating and maintaining communities (Muniz & O'Guinn, 2000) and Internet technologies greatly facilitate this form of communication and community building among travelers. Importantly, the new scarcities of time and trust require new tourism consumers to rely heavily on word of mouth and the expert opinions of like-minded others. New travel-oriented communities are brand communities or communities of interest and are imagined, involve limited liability, and focus on a specific consumption practice (Muniz & O'Guinn, 2000; Wang, et al., 2002).

A wealth of information creates a poverty of attention (Lewis & Bridger, 2000). New tourism consumers try to cope with this problem by scanning information depending on personal relevance and have become very capable of ignoring nonrelevant advertising. They are, therefore, much more active in their travel information search than old consumers, who were largely passive information recipients. Attention is increasingly reserved for marketers who have asked for permission and have established a long-term relationship with the consumer (Godin, 1999). In return for his/her valuable attention, the new tourism consumer expects special benefits such as extremely personalized services. Attention peaks when these travelers reach a psychologically balanced state of mind, a so-called "flow" experience (Feather, 2000). In order to reach flow, new tourism consumers are increasingly searching for personalized, emotional, and intriguing experiences through which they can learn about new travel and tourism products. Therefore, in the world of the new tourism consumer, the focus shifts from product attributes to consumption experiences.

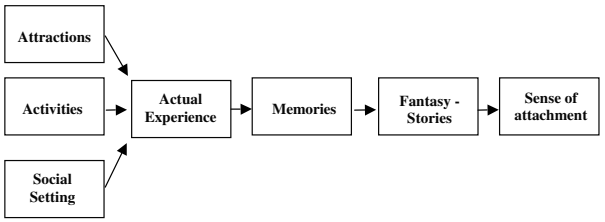


Figure 3: Sequence of travel experience.

Trend #5. Emergence of Experience as the Foundation for Defining Tourism Products

It has long been recognized that travel is an experience and tourism is a key part of the "experience industry" (Pine, Gilmore, & Pine, 1999). However, the role of experience in consumption (including pre-, during, and postpurchase) is only now being considered as one of the foundations for effective marketing. Recent efforts have shown that the experiential aspects of products and services provide the starting point for effective marketing (O'Sullivan & Spangler, 1998; Pine et al., 1999; Schmitt, 1999). This research indicates that experiences are personal "events" that engage the individual in a meaningful way. As shown in Figure 3, the core element of travel experiences is the travel activity, whereas the tourism industry plays the part of an experience "facilitator"; importantly, the setting (social or personal) in which activities occur contributes substantially to the nature of the experience. It is suggested that although the experiential aspects of travel are the foundation, the memories that are stored as a result of these experiences are the key to attracting new visitors as well as retaining current ones. Furthermore, it is suggested that stories—the mechanisms for communicating experiences through word of mouth or as "documentaries" of experiences (through articles, film, etc.) provide the path through which the tourism industry can build and extend markets.

Schmitt (1999) and others have argued that the new consumer evaluates products more on their experiential aspects than on "objective" features such as price and availability and that experiential marketing should focus on the experiential aspects that make the consumption of the product most compelling—that is, the five senses. Effective experiential marketing is sensory and affective. It approaches consumption as a holistic experience and acknowledges that consumers can be either rational or emotional or both at the same time. Whereas traditional marketing is based on consumer behavior, product features, benefits, and quantifiable market segments, experiential marketing is driven by an understanding of consumer experiences and the need for personalization. New consumers require advertising that is entertaining, stimulating, and at the same time informative. Brands are no longer seen as mere identifiers but become themselves sources of experiences by evoking sensory, affective, creative, and lifestyle-related associations (Schmitt, 1999). Thus, experiential marketing blurs the border between advertising, purchase, and use as it attempts to create a unique shopping experience and lets the new consumer anticipate what the consumption experience will be like.

FUTURE BEHAVIOR IN TRAVEL

The following briefly summarizes some expectations for the future role of the Internet in travel and tourism.

Travel will continue to be one of the most popular online interests to consumers. This trend will increase in magnitude as travel providers create more effective means with which to communicate the nature of their offerings.

The Internet and alternate access devices are increasing the number of electronic connections between customers and the tourism industry. These new technologies will continue to provide an environment for creating relationships, allowing consumers to access information more efficiently, conducting transactions, and interacting electronically with businesses and suppliers. Examples of emerging technology in the travel industry include travel recommendation systems, travel guidance systems and virtual reality.

The changes in demographic profiles of Internet users over the past decade suggest that the evolving Internet and related systems will ultimately be adopted by the large majority of the traveling public and, therefore, the Internet will be considered the primary source for travel information.

The demands of travelers, and in particular the purchase process(es) they use, will continue to evolve as consumers of travel products gain more experience and confidence in product purchasing over the Internet. Importantly, conversations among travelers (through travel clubs, virtual communities, etc.) will continue to grow and will increasingly be mediated through Internet technologies.

Experience- and emotion-oriented Internet communications will grow in importance as human-centric computing and emotionally intelligent interfaces are offered on the Internet. These interfaces/systems will incorporate a variety of interpreted information, enabling the systems to recognize the information needs of the user within an emotional-psychological need context, in order to provide supportive interactions and suggestions.

The trends identified above set the stage for an interesting and challenging future for the travel and tourism industry. Following from Naisbitt (1994), the "global paradox" for travel organizations lies in having to compete at the local level for individual travelers but also simultaneously at the national and international levels. The innovative power of the Internet provides stimulating input for new organizational strategies but at the same time constrains the ability of current organizations to adjust to the "new realities." The rich informational environment the Internet provides and the availability of and access to an "infinite" number of "experiential settings" empower consumers in a variety of ways. The challenge for tourism organizations is to set stages for experience creation throughout their organizational structures and to actively involve all employees in the design and marketing of experiences so that the full benefits of the Internet can be realized.

GLOSSARY

Brochureware A term used to refer to Web sites or Web pages created by taking printed tourism brochures and directly transferring their contents to digital environments. Web sites designed as brochureware represent the simplest form of Web design as they display information in a static way and provide only limited navigation and communication functions.

Central reservation system (CRS) A system that provides access to information about airline, hotel, or rental company inventories and is used for sales, marketing, and ticketing purposes. The elements of a CRS include a central processing unit, a central database, and a communications network that links information providers and users to the central information storage and processing system.

Click stream A recorded path of the pages a user requests in navigating through one or more Web sites. Click stream information can help Web site owners and advertisers understand how visitors use a site. More specifically, it provides insights with respect to how the site was found, how much time was spent on the site, and what specific pages were accessed.

Customer relationship management (CRM) A set of business principles employed with the aim of strengthening a tourism organization's relationships with its clients, optimizing customer service levels, and obtaining customer information that can be used for marketing purposes.

Destination-management system (DMS) The information technology infrastructure used by destination marketing organizations to support a wide range of promotion, sales, and advertising efforts. Typical elements of such systems are technologies to design and produce printed materials, tourist information center services (including information and reservation systems), call center services, kiosks, database marketing applications, project/event management software, and marketing research applications.

Destination marketing organization (DMO) is an organization with responsibility for marketing a specific tourism destination to the travel trade and to individual travelers.

Distribution channel A set of interdependent organizations, such as tour operators or travel agents, involved in the process of making tourism products or services available to consumers.

En route information Information obtained while traveling, as opposed to information collected before or after the trip; mainly used for navigational purposes or to support short-term decision making during the actual vacation.

Flow A seamless, intrinsically enjoyable, self-reinforcing, and captivating psychological experience occurring when there is an optimal match between the challenge at hand and one's skills.

Global distribution system (GDS) A system that links several central reservation systems, thus providing a much more global coverage than individual computerized reservation systems.

Human-centric computing The process of designing, developing, and implementing information technology that reflects the needs and lifestyles of its human users.

Information asymmetry Also referred to as information gap; a condition in which at least some relevant information is known to some but not all parties involved. Information asymmetry causes markets to become inefficient, because not all players have equal access to information needed for decision-making processes.

Intermediary A third party organization such as a travel agent, tour operator, incoming agent, hotel chain, or CRS/GDS provider that facilitates information transfer and/or transactions between the primary suppliers and consumers of tourism products and services.

Telepresence A mental state in which a user feels physically present within a remote, computer-mediated environment.

Travel and tourism industry The individuals and organizations that are involved in the production and distribution of travel and tourism-related products.

Travel decision support system (TDSS) An information system designed to simplify the travel decision-making process and support consumers in the various steps involved in planning trips.

Travel destination The end-point of a trip or trip segment. The term is also used to describe a geographically or perceptionally defined area of particular interest to tourists and includes all attractions and other tourism establishments that exist within its boundaries.

Tourism experience The sum of sensory, cognitive, and emotional inputs derived from the activities pursued during a vacation.

Virtual travel community A community facilitated by computer-mediated communication that allows its members to conduct various types of travel-related tasks, such as obtaining travel information, maintaining connections, finding travel companions, or simply having fun by telling each other interesting stories about their travel experiences.

Virtual tour A tool that enables potential consumers of tourism products to explore and immerse themselves within an interactive Web environment in order to gain the needed experiential information about a destination or tourism establishment.

CROSS REFERENCES

See *Customer Relationship Management on the Web; Internet Literacy; Online Communities; Web Site Design.*

REFERENCES

Brownlie, D., Saren, M., Whittinton, R., & Wensley, R. (1994). The new marketing myopia: Critical perspectives on theory and research in marketing–Introduction. *European Journal of Marketing, 28*(3), 6–12.

Cho, Y., & Fesenmaier, D. R. (2001). A new paradigm for tourism and electronic commerce: Experience marketing using the virtual tour. In E. Laws & D. Buhalis (Eds.), *Tourism distribution channels: Practices, issues and transformations* (pp. 351–370). New York: CAB International.

Connolly, D., & Olson, M. (1999). *Hospitality technology in the new millennium: Findings of the IH&RA think-tanks on technology*. Paris: International Hotel & Restaurant Association.

Connolly, D., Olson, M., & Moore, R. G. (1998, August). The Internet as a distribution channel. *Cornell Hotel and Restaurant Administrative Quarterly,* 42–54.

Connolly, D., & Sigala, M. (2001). Major trends and IT issues facing the hospitality industry in the new economy: A review of the 5th Annual Pan-European Hospitality Technology Exhibition and Conference. *International Journal of Tourism Research, 3,* 325–327.

Cronin, M. J. (1995). *Doing business on the Internet.* New York: Van Nostrand Reinhold.

Day, G. S. (1998). Organizing for interactivity. *Journal of Interactive Marketing, 12*(1), 47–53.

Delgado, J., & Davidson, R. (2002). Knowledge bases and user profiling in travel and hospitality recommender systems. In K. Wöber, A. J. Frew, & M. Hitz (Eds.), *Information and Communication Technologies in Tourism 2002, Proceedings of the Annual Conference of the International Federation of Information Technology and Tourism* (pp. 1–16). Vienna, Austria: Springer-Verlag.

Dellaert, B. G. C., Ettema, D. F., & Lindh, C. (1998). Multi-faceted tourist travel decisions: A constraints-based conceptual framework to describe tourists' sequential choices of travel components. *Tourism Management, 19*(4), 313–320.

Dertouzos, M. (2001). *The unfinished revolution*. New York: HarperCollins Publishers.

eMarketer (1999). *eAdvertising report volume II*. Retrieved December 23, 2002, from http://www.emarketer.com

Eriksson, O. (2002). Location based destination information for the mobile tourist. In K. Wöber, A. J. Frew, & M. Hitz (Eds.), *Information and Communication Technologies in Tourism 2002, Proceedings of the Annual Conference of the International Federation of Information Technology and Tourism* (pp. 255−264). Vienna, Austria: Springer-Verlag.

Evans, P., & Wurster, T. S. (1999). *Blown to bits*. Boston: Harvard Business School Press.

Feather, F. (2000). *FutureConsumer.Com*. Toronto: Warwick Publishing.

FIND/SVP (1994–1997). *American Internet user surveys developed by the Cyber Dialogue.* New York: FIND/SVP.

Forrester Research (1999). *Forrester report: Dynamic trade voyage*. Retrieved December 23, 2002, from http://www.forrester.com/Voyage/PDF/0,2236,4823,00.pdf

Gartrell, R. (1988). *Destination marketing for convention and visitor bureaus*. Dubuque, IA: Kendall/Hunt Publishing Company.

Godin, S. (1999). *Permission marketing*. New York: Simon & Schuster.

Goeldner, C. R., & Ritchie, J. R. B. (2002). *Tourism: Principles, practices, philosophies,* (9th ed.). New York: Wiley.

Good, N., Schafer, J. B., Konstan, J., Borchers, A., Sarwar, B., Herlocker, J., & Riedl, J. (1999). Combining collaborative filtering with personal agents for better recommendations. In J. Hendler, D. Subramanian, R. Uthurusamy, & B. Hayes-Roth (Eds.), *Proceedings of the Sixteenth National Conference on Artificial*

Intelligence and Eleventh Conference on Innovative Applications of Artificial Intelligence (pp. 439–446). Menlo Park, CA: AAAI Press.

Gretzel, U., Yuan, Y., & Fesenmaier, D. R. (2000). *White paper on advertising and information technology in tourism*. Champaign, IL: National Laboratory for Tourism and eCommerce, University of Illinois.

Hanson, W. (2000). *Principles of Internet marketing*. Cincinnati, OH: South-Western College Publishing.

Hardie, M., Bluestein, W., McKnight, J., & Davis, K. (1997). *The Forrester report: Entertainment & technology strategies*. Retrieved December 20, 2002, from http://www.forrester.com

Hibbard, J. (1997). Straight line to relevant data. *Information Week, 657,* 21–25.

Hoffman, D., & Novak, T. (1996, July). Marketing in hypermedia computer-mediated environments: Conceptual foundations. *Journal of Marketing, 60,* 50–68.

Hoffman, D., Novak, T., & Chatterjee, P. (1995). *Commercial scenarios for the Web: Opportunities and challenges*. Retrieved December 20, 2002, from http://www.ascusc.org/jcmc/vol1/issue3/hoffman.html

Hruschka, H., & Mazanec, J. (1990). Computer-assisted travel counseling. *Annals of Tourism Research, 17,* 208–227.

Hwang, Y-H., & Fesenmaier, D. R. (2001). Collaborative filtering: Strategies for travel destination bundling. In P. Sheldon, K. Wöber, & D. R. Fesenmaier (Eds.), *Information and Communication Technologies in Tourism 2001, Proceedings of the Annual Conference of the International Federation of Information Technology and Tourism* (pp. 167–175). Vienna, Austria: Springer-Verlag.

Information Society Technologies Advisory Group (2001). *Scenarios for ambient intelligence in 2010*. Retrieved December 20, 2002, from ftp://ftp.cordis.lu/pub/ist/docs/istagscenarios2010.pdf

Inkpen, G. (1998). *Information technology for travel and tourism.* Singapore: Addison Wesley Longman Limited.

Jeng, J-M., & Fesenmaier, D. R. (2002). Conceptualizing the travel decision-making hierarchy: A review of recent developments. *Tourism Analysis, 7*(1), 15–32.

Jupiter Communications (1997). *Travel and interactive technology: A five year outlook*. Washington, DC: Travel Industry Association of America.

Klein, S., & Langenohl, T. J. (1994). Co-ordination mechanisms and systems architectures in electronic market systems. In W. Schertler, B. Schmid, A. M. Tjoa, & H. Werthner (Eds.), *Information and Communication Technologies in Tourism 1994, Proceedings of the Annual Conference of the International Federation of Information Technology and Tourism* (pp. 262–270). Vienna, Austria: Springer-Verlag.

Kotler, P., Bowen, J., & Makens, J. (1999). *Marketing for hospitality and tourism* (2nd ed.). Upper Saddle River, NJ: Prentice–Hall, Inc.

Lewis, D., & Bridger, D. (2000). *The soul of the new consumer.* London: Nicholas Brealey Publishing.

Loban, S. (1997). A framework for computer-assisted travel counseling. *Annals of Tourism Research, 24*(4), 813–831.

Mayros, V., & Werner, D. M. (1982). *Marketing information systems: Design and applications for marketers*. Radnor, PA: Chilton Book Company.

McGuffie, J. (1994). CRS development in the hotel sector. *EIU Travel and Tourism Analyst, 2,* 53–68.

Miller, M. J. (2002, March 12). Living history—Retracing the evolution of the PC and PC Magazine. *PC Magazine,* 137–159.

Mitsche, N. (2001). Personalized traveling counseling system: Providing decision support features for travelers. In P. Sheldon, K. Wöber, & D. R. Fesenmaier (Eds.), *Information and Communication Technologies in Tourism 2001, Proceedings of the Annual Conference of the International Federation of Information Technology and Tourism* (pp. 160–166). Vienna, Austria: Springer-Verlag.

Morrison, A. M., Taylor, S., Morrison, A. J., & Morrison, D. (1999). Marketing small hotels on the World Wide Web. *Information Technology and Tourism, 2*(2), 35–44.

Moutinho, L. (1987). Consumer behavior in tourism. *European Journal of Marketing, 21,* 2–44.

Muniz, A. M., & O'Guinn, T. C. (2000, March). Brand community. *Journal of Consumer Research, 27,* 227–235.

Naisbitt, J. (1994). *Global paradox*. New York: Avon.

Newell, F. (2000). *Loyalty.com*. New York: McGraw–Hill.

Norman, D. A. (1999). *The invisible computer: Why good products can fail, the personal computer is so complex and information appliances are the solution*. Cambridge, MA: MIT Press.

O'Sullivan, E. L., & Spangler, K. J. (1998). *Experience marketing*. State College, PA: Venture Publishing.

Oertel, B., Steinmüller, K., & Kuom, M. (2002). Mobile multimedia services for tourism. In K. Wöber, A. J. Frew, & M. Hitz (Eds.), *Information and Communication Technologies in Tourism 2002, Proceedings of the Annual Conference of the International Federation of Information Technology and Tourism* (pp. 265–274). Vienna, Austria: Springer-Verlag.

Parsons, A., Zeisser, M., & Waitman, R. (1998). Organizing today for the digital marketing of tomorrow. *Journal of Interactive Marketing, 12*(1), 31–46.

Pine, B. J., Gilmore, J. H., & Pine, B. J., II, (1999). *The experience economy*. Boston, MA: Harvard Business School Press.

Pitkow, J. (1997). *The WWW user population: Emerging trends*. GVU Center, Atlanta: Georgia Institute of Technology.

Poon, A. (1993). *Tourism, technology and competitive strategies*. Wallingford, UK: CAB International.

Poslad, S., Laamanen, H., Malaka, R., Nick, A., Buckle, P., & Zipf, A. (2001). CRUMPET: Creation of user-friendly mobile services personalized for tourism. In *Proceedings of 3G 2001*. London: Institution of Electrical Engineers. Retrieved December 22, 2002, from http://www.eml.villa-bosch.de/english/homes/zipf/3g-crumpet2001.pdf

Rheingold, H. (1994). A slice of life in my virtual community. In L. M. Harasim (Ed.), *Global Networks: Computers and International Communications* (pp. 57–80). Cambridge, MA: MIT Press.

Ricci, F., & Werthner, H. (2001). Cased-based destination recommendations over an XML data repository.

In P. Sheldon, K. Wöber, & D. R. Fesenmaier (Eds.), *Information and Communication Technologies in Tourism 2001, Proceedings of the Annual Conference of the International Federation of Information Technology and Tourism* (pp. 150–159). Vienna, Austria: Springer-Verlag.

Schmidt-Belz, B., Makelainen, M., Nick, A., & Poslad, S. (2002). Intelligent brokering of tourism services for mobile users. In K. Wöber, A. J. Frew, & M. Hitz (Eds.), *Information and Communication Technologies in Tourism 2002, Proceedings of the Annual Conference of the International Federation of Information Technology and Tourism* (pp. 265–274). Vienna, Austria: Springer-Verlag.

Schmitt, B. H. (1999). *Experiential marketing: How to get customers to sense, feel, think, act, and relate to your company and crands*. New York: Free Press.

Schwartz, E. I. (1998). *Webonomics: Nine essential principles for growing your business on the World Wide Web*. New York: Broadway Books.

Sheldon, P. J. (1997). *Tourism information technology*. Wallingford, UK: CAB International.

Sterne, J. (1999). *World Wide Web marketing*. New York: Wiley.

Travel Industry Association of America (1998). *Technology and travel 1998: Executive report*. Washington, DC: Travel Industry Association of America.

Travel Industry Association of America (2000). *Technology and travel 2000: Executive report*. Washington, DC: Travel Industry Association of America.

Travel Industry Association of America (2002a). *Technology and travel 2002: Executive report*. Washington, DC: Travel Industry Association of America.

Travel Industry Association of America (2002b). *U.S. travel industry performance update #10*. Washington, DC: Travel Industry Association of America.

University of California, Los Angeles (2001). *The UCLA Internet report 2001—Surveying the digital future*. Los Angeles: Center for Communication Policy.

Vanhof, K., & Molderez, I. (1994). An advice system for travel agents. In W. Schertler et al. (Eds.), *Information and Communication Technologies in Tourism 1994, Proceedings of the Annual Conference of the International Federation of Information Technology and Tourism* (pp. 126–132). Vienna, Austria: Springer-Verlag.

Vogt, C. A., & Fesenmaier, D. R. (1998). Expanding the functional information search model. *Annals of Tourism Research, 25*(3), 551–578.

Wang, Y., & Fesenmaier, D. R. (2002). *Assessing Web marketing strategies: Approaches, issues, and implications: A report on the results of national survey of city and county tourism organizations in the United States of America*. Champaign, IL: National Laboratory for Tourism and eCommerce, University of Illinois.

Wang, Y., Yu, Q., & Fesenmaier, D. R. (2002). Defining the virtual tourist community: Implications for tourism marketing. *Tourism Management, 23,* 407–417.

Werthner, H., & Klein, S. (1999). *Information technology and tourism——A challenging relationship*. Vienna, Austria: Springer-Verlag.

Wind, Y., Mahajan, R., & Gunther, R. (2002). *Convergence marketing*. Upper Saddle River, NJ: Prentice Hall.

Windham, L., & Orton, K. (2000). *The soul of the new consumer.* New York: Allworth Press.

Wiseman, C. (1985). *Strategy and computers: Information systems as competitive weapons*. Homewood, IL: Dow Jones–Irwin.

Woodside, A. G., & MacDonald, R. (1994). General system framework of customer choice processes of tourism services. In V. Gasser & K. Weiermair (Eds.), *Spoilt for choice. Decision making processes and preference change of tourists: Intertemporal and intercountry perspectives* (pp. 30–59). Thaur, Austria: Kulturverlag.

World Tourism Organization (2001). *Tourism statistics 2000*. Retrieved January 20, 2003, from http://www.world-tourism.org/market_research/facts&figures/menu.htm

World Tourism Organization Business Council (1999, September). *Marketing tourism destinations online*. Madrid, Spain: World Tourism Organization Business Council.

World Travel and Tourism Council (2002). *Year 2002—End of year update, TSA research summary and highlights*. Brussels: World Travel and Tourism Council.

Yuan, Y., & Fesenmaier, D. R. (2000). Preparing for the new economy: The use of the Internet and intranet in American convention and visitor bureaus. *Information Technology and Tourism, 3*(2), 71–86.

Zeff, R., & Aronson, B. (1999). *Advertising on the Internet* (2nd ed.). New York: Wiley.

U

Universally Accessible Web Resources: Designing for People with Disabilities

Jon Gunderson, *University of Illinois at Urbana–Champaign*

INTRODUCTION

Tim Berners Lee developed the first HTML (hypertext markup language) Web browser/editor in 1990 to enable scientists at the CERN particle physics lab in Switzerland to share electronic documents on a wide range of computing systems. At the heart of the design of HTML technologies is the concept of interoperability, the ability of providing and receiving electronic documents using public standards for creating, serving, and viewing the information on a wide variety of computing equipment. In the beginning the focus was on the information. Users typically had a wide range of choices and control over the rendering of Web documents. Authors were not very interested in controlling the rendering of HTML and indeed HTML has limited features for absolute control over rendering.

MYTHS OF THE WEB

As the Web was commercialized through the introduction of graphical browsers (NCSA Mosaic, Netscape Navigator, and Microsoft Internet Explorer) in the mid- to late 1990s there was a fundamental change in the relationship between the control authors had over the rendering style of Web resources and the users' ability to control the rendering. There are many reasons for this shift, but the result is that the vast majority of Web authors developing commercial content primarily think of the Web as a graphical medium. At the same time the most popular commercial browser developers provided users with fewer and fewer options for adjusting the rendering of Web resources to a point where most users today do not know they have any control over the rendering of Web content, and this reinforces the beliefs of the Web as a graphical medium in which users have no control over rendering. This has lead to the design of inaccessible Web resources that increasingly only support graphical renderings and the use of pointing devices (e.g., the mouse) for interacting with dynamic content. An example of this narrowing view of interoperability is many developers requiring their Web pages appear visually the same in both Netscape Navigator 4.7 and Internet Explorer 4.0+ even though Netscape Navigator 4.7 is an outdated technology that does not conform to HTML 4.0 or 3.2 specifications (CITA Surveys, 2001a, 2001b). This approach leads developers to use images and complex table layouts for styling and positioning text and images, giving users little opportunity to access the content in non-graphical renderings of text, Braille, or speech.

DIGITAL DIVIDE

The divide between people with visual impairments and able-bodied Web users was investigated by Pernice-Coyne and Nielsen (2001). They found that people who use screen magnification technologies could only complete Web tasks about 21% of the time and people using speech output about 12.5% of the time. When compared to the able-bodied control group performance of completing tasks about 78% of the time, it is clear that current Web design is creating tremendous barriers to people with

disabilities. When the visually impaired and blind did complete tasks they took about twice as much time and visited twice as many Web resources as the control group. This indicates that Web resources are not providing information on the structure or organization of their Web pages so that they can be used by people with disabilities to efficiently identify and find the information they seek.

It is estimated that in 1997 approximately 48 million Americans over the age of 15 years old have some type of disability and that about 17 million identified themselves as having a severe disability (US Census Bureau, 2001). As people age the percentage with disabilities increases from 1.6% for people between the ages of 15 and 24 years old, to 5% for those between the ages of 21 and 64, and then triples to 17% for those over the age of 65. So a major part of the financial argument for designing universally accessible Web resources is designing resources that deal with an increasingly aging population and the economic power and productive capacities they bring to our nation's economy. Kay (2000) found this barrier in the use of computer technology by people with disabilities, who own computers at half the rate as the general population and use the Internet about a fourth of the time. There are many factors that influence computer ownership and Internet use. Probably one of the most critical factors is how the technology is designed to be inclusive of the needs of people with disabilities. Before concrete ramps and curb cuts were built into the physical structures of our society few places were accessible to people with disabilities who were thus invisible to much of the general public. In the same way, electronic ramps and curb cuts need to be built into our electronic infrastructure before we will see the wide-spread presence of people with disabilities on the Internet.

ALTERNATIVE VIEWS OF THE WEB

For the Web to become more accessible to people with disabilities Web authors need to understand that people will be viewing their resources through many different renderings, including graphical, text, and speech. This section outlines the technologies people with disabilities use to access the Web, including the rendering options of popular Web browsers and specialized technologies designed specifically for people with disabilities.

The W3C User Agent Accessibility Guidelines (Jacobs, Gunderson, and Hansen, 2002) outline the types of features browsers and multimedia players need to provide in order for people with disabilities to be able to access Web content. One of the primary requirements is the ability to support the keyboard. People with many types of disabilities for various reasons can only use the keyboard to control the browser. Therefore, functions not available through keyboard commands will not be available to people with disabilities. People with disabilities also need to be able to select what types of content they want to view. For example, people who cannot see images benefit more from text descriptions of the images. They would configure the browser to render the text description of an image in place of the graphical image. Other types of control include the ability to control the styling of text font characteristics and the foreground and background colors where the text is rendered. People with visual impairments often need to use sans serif fonts, larger text size, and specific color combinations to make the text readable. Automatic behaviors supported by many graphical browsers, like authors automatically generating new windows, can be disorienting because the user is not expecting a new window to open when they follow a link. Usually the new window is given focus. When the user tries to use the back function of the browser to reorient themselves to the previous page the page does not change, since the new window does not inherit the history of the previous window. This is a usability problem for all users, but it has an increased impact on people with disabilities since they often do not have information about the choices available to them and, as well, are not oriented to the new window being open.

People who are blind cannot use the computer screen at all and use synthetic speech and refreshable Braille displays to access Web content. Speech navigation and browsing is much different than graphical browsing since the user is only able to view the content linearly. When using speech it is important to provide markup that allows users to skip to important structures like headers, navigation bars, and links. Otherwise, users need to read the entire document to understand the information available.

KEYBOARD SUPPORT

Applications written for Microsoft Windows typically have very good keyboard support in contrast to applications written for Apple Macintosh and the various flavors of the UNIX X-Windows systems. This generalization is also applicable to browsers. Browsers like Netscape Navigator (http://www.netscape.com), Internet Explorer (http://www.microsoft.com/ie), and Operasoft Opera (http://www.opera.com) all have keyboard shortcuts for the following functions:

Next link or form control
Previous link or form control
Select link
Move focus to next frame
Reload content (Refresh)
Stop loading content
Move to previous resource in history
Move to next resource in history
Show next page of content
Show previous page of content
Next frame
Menu bar options
Navigation and setting of controls in dialog boxes

Opera provides additional keyboard commands that allow the user to navigate the structural elements of Web resources. This includes individual functions to navigate by headers (H1–H6), form controls, and HTML element-by-element. This type of function allows users to more efficiently identify the main topics of a Web page, since they do not need to rely on the styling the author used for the header to identify the text as a major topic. On pages with a large number of links it can be rather tedious to navigate to a specific link using the simple next link function available in Internet Explorer and Netscape

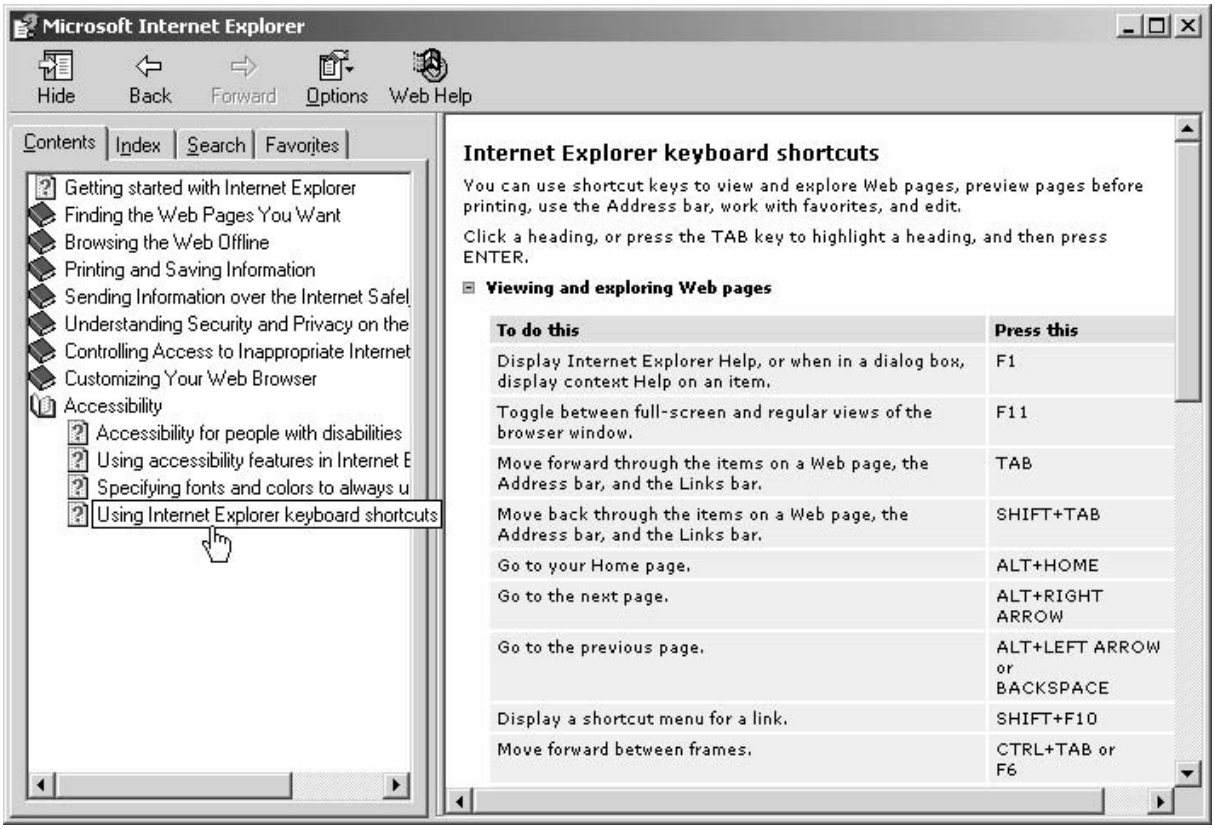


Figure 1: Help files for Internet Explorer keyboard shortcuts.

Navigator. Opera allows the user to navigate past large numbers of links (if headers are used properly by the author) to the header closest to the link they want to select and then use the next link function from this position. Opera has a second function for navigating to links whereby the user can use keyboard commands to view the list of links in a document or individual links, using letter keys to move to those links that start with that letter. This type of function helps people with minimal range of motion use their physical ability more efficiently and people with visual impairments to have more options for searching and selecting a link. The keyboard shortcuts for a browser or multimedia player can typically be found in the help system (Figure 1).

ACCESS TO TEXT DESCRIPTIONS

One of the most important configuration options is the ability to render text descriptions for images. HTML has two attributes of the IMG element that can be used to provide text descriptions, the ALT attribute for short descriptions and the LONGDESC attribute for providing a link to a longer description. Most graphical browsers render ALT text content in place of an image when the browser is configured to not render images, but the quality of the rendering varies considerably among current browser technology. One of the major issues with rendering text descriptions is the difference in space required to render the text descriptions. Often text descriptions require more graphical space than the original image, requiring a re-flow of content to accommodate the text description. When images are used for spacing and positioning this can often create a distorted rendering of text, making it more difficult for the user to understand the content relationships.

Currently the major graphical browsers Opera 6.05, Internet Explorer 6.0, and Netscape 7.0 do not fully support access to the text descriptions for all images, only a subset or under special conditions. The HTML IMG element is the most popular way authors include images in Web pages, but other elements including AREA and INPUT can have ALT attribute content. The IMG element includes an ALT attribute and a LONGDESC attribute for associating text descriptions with images. Table 1 shows the capabilities of various browsers in rendering ALT text descriptions.

Figures 2, 3, and 4 show the ALT text rendering of the same Web page for Opera 6.1, Internet Explorer 6.0, and Netscape Navigator 7.0, respectively. Opera renders the ALT text and has extensive styling capabilities for the ALT text. Internet Explorer for Windows renders that ALT text and limits the ability of the user to style the ALT text. Netscape Navigator does not render ALT text when rendering of images are turned off.

The text content of the ALT attribute is designed to provide a short text description of an image. The LONGDESC attribute provides a URI to a Web resource that will provide a more detailed description of the image. For example, if the image was a chart of what flavors of ice cream people prefer at a certain ice cream store, the LONGDESC could point to a Web page with a text table representation of the ice cream preferences. Opera has a very good implementation of rendering ALT text, since it provides the user with extensive control over styling the ALT text. Other

Table 1 Browser Capabilities of Rendering ALT Text and Providing a Link to LONGDESC URI

Browser	Operating System	Render ALT Text for IMG Element	Render ALT Text for AREA Element	User Styling of ALT Text	Link to LONGDESC URI
Opera 6.01	Windows 98/2000/XP, Macintosh OS9, and UNIX	Yes	No	Yes	No
Internet Explorer 6.0	Windows 98/2000/XP	Yes, except when scripts dynamically change the image source attribute	No	Limited	No
Internet Explorer 5.1	Macintosh OS9	Yes, unless images are cached	No	No	No
Netscape 7.0	Windows 98/2000/XP, Macintosh OS9, and UNIX	No	No	No	Yes, through Context menu that is only accessible with mouse commands

implementations, including Internet Explorer, provide varying levels of access to ALT text, ranging to no access in the case of Netscape 6.2. Access to ALT text has much better implementation than access to the LONGDESC attribute. It is ironic that the one browser providing access to the LONGDESC URI, Netscape 6.2, does not provide access to the ALT text description. Table 1 shows that no browser provides complete access to text descriptions for the IMG element, limiting the types of content users will have access to when using these browsers.

The <INPUT> element of type "image" and <AREA> element that defines the clickable areas on an image MAP also have ALT attributes. Currently none of the major graphical browsers support the in-content rendering of ALT text associated with the AREA element, some provide access through display as a tool tip (e.g., a pop text description of an element that appears when the pointing device hovers over an element rendering) and many assistive technologies like screen can read the value of the ALT text. The lack of support makes the links of the image MAP elements inaccessible to many people with disabilities who have poor vision and do not use assistive technologies. Authors should always provide redundant text links for both server-side and client-side image MAPs.

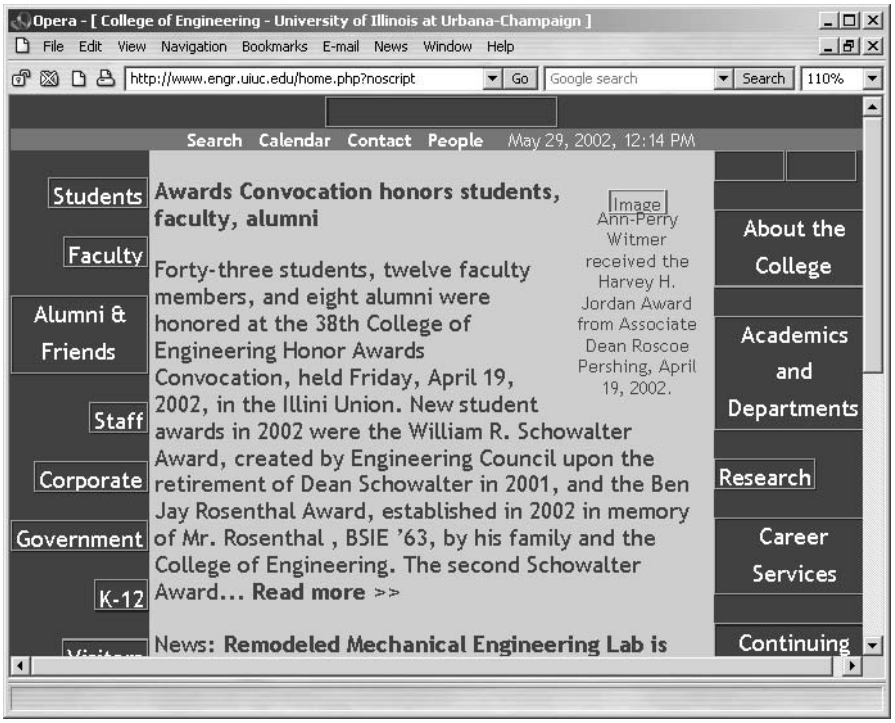


Figure 2: View of ALT text rendering in Opera.

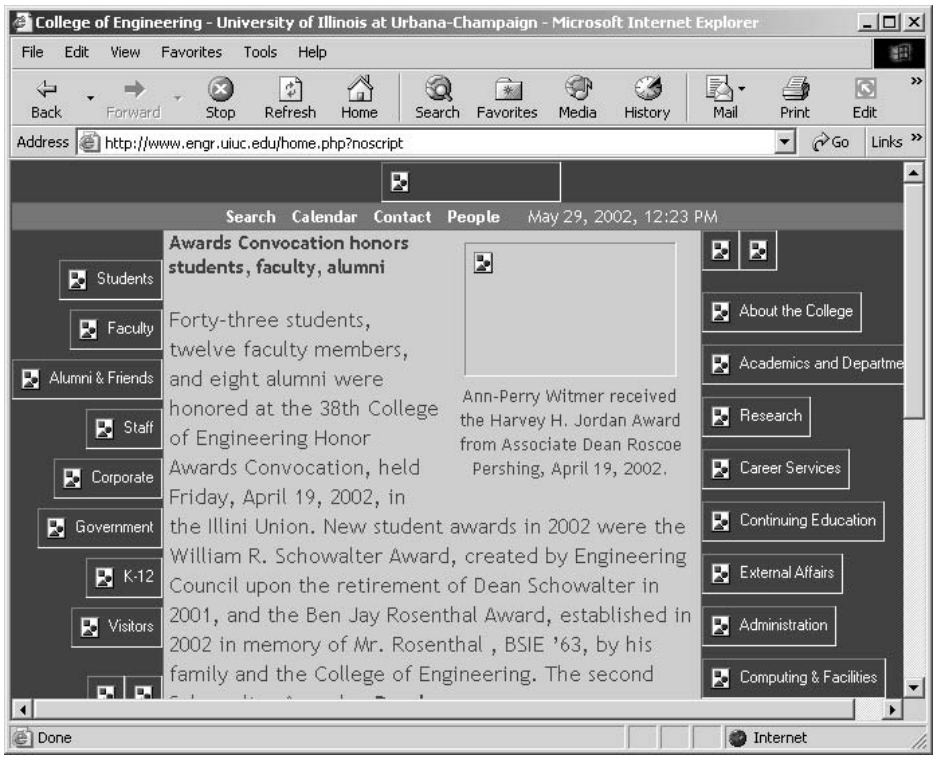


Figure 3: View of ALT text rendering in Internet Explorer.

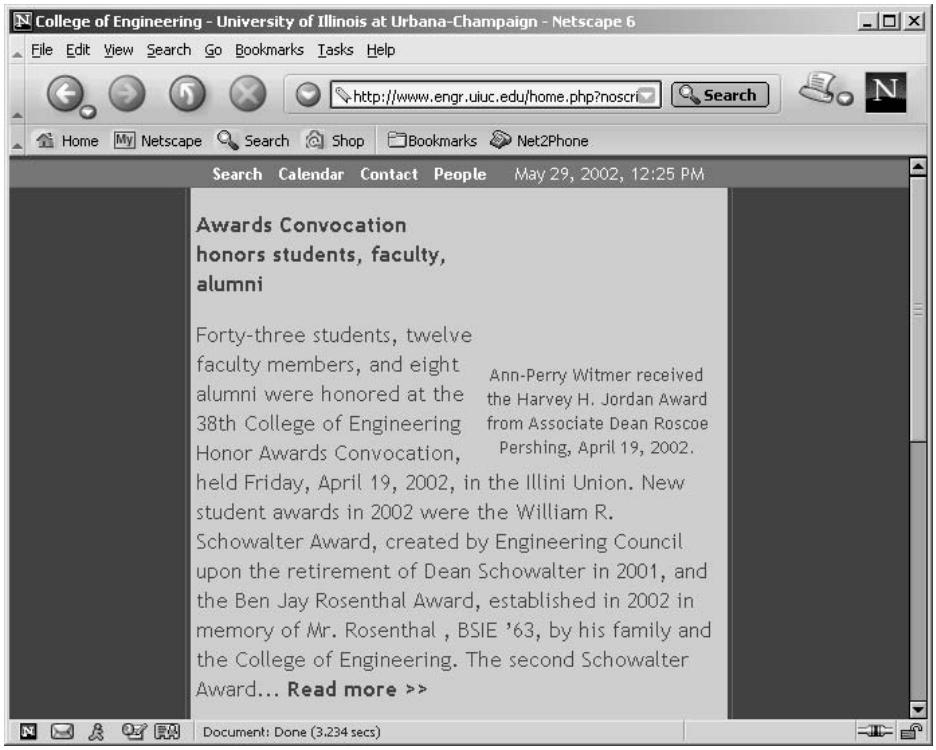


Figure 4: View of ALT text rendering in Netscape Navigator.

Table 2 Browser Capabilities of Overriding Author Styles with User Style Sheets

Browser	Operating System	Ignore Author Font Sizes, Font Style and Colors	Adjust Default Font Sizes, Font Style and Colors	Ignore Author Style Sheets	Add User Style Sheet
Opera 6.01	Windows 98/2000/XP, Macintosh OS9, and UNIX	Yes	Yes	Yes	Yes
Internet Explorer 6.0	Windows 98/2000/XP	Yes	Yes, limited font size control except through style sheets	Limited to fonts and colors	Yes
Internet Explorer 5.5	Macintosh OS9	Yes	Yes	Yes	Yes, but only in cascade with author style sheets (i.e., cannot use user style sheets when author style sheets are turned on)
Netscape 6.2	Windows 98/2000/XP, Macintosh OS9, and UNIX	Yes	Yes	Limited to fonts and colors	No

USER STYLING OF TEXT

People with visual impairments and learning disabilities that affect reading need to control the font characteristics, font size, and foreground and background colors of text. In early graphical browsers, like NCSA Mosaic, this was a built-in feature and the user could completely configure the default style sheet used for rendering HTML. Current browsers vary widely in their ability to allow the user to control the rendering of text. Table 2 shows the capabilities of several popular browsers in allowing the user to control the rendering of text.

The W3C Cascading Style Sheet (Lie & Bos, 1997) technology was designed to address many of the author and user styling issues of separating the structure of a Web resource and the styling for a particular rendering. The advantage to developers in using style sheet technology is that a single style sheet can be used to control the rendering of any number of Web pages, making it easier for webmasters to change the look and feel of their Web site without having to individually edit pages or elements. One of the more powerful aspects of the CSS specifications is user style sheets. The W3C realized that users need control over rendering and the specification includes the concept of user style sheets overriding author-supplied style sheets. Opera has actually implemented the concept of user and author styling and provides a very concrete interface for users to select author and user styling preferences (Figure 5). Opera also provides a one-step command (clickable icon or single key press) for the user to switch between author styling and user styling of Web content. This is a very useful feature not only for users, but also Web developers. Authors can easily switch between their designs and a high-contrast styling that might be used by someone with a visual impairment. The high-contrast setting helps them to determine whether their Web design will work for someone with a severe visual impairment or using portable technologies like a PDA or speech browser that does not have the same rendering characteristics as a graphical browser. Microsoft Internet Explorer implements user style sheets, but does not allow the user to completely ignore style sheets supplied by the author. Table 2 shows the features available to users to control the author-supplied styles and to apply their own style sheets in various browsers.

SPEECH BROWSING

Speech browsing is a fundamentally different way of accessing Web information. In a graphical rendering the author often uses spatial relationships to group information on the screen and the users passively scan the screen to identify the grouping of information. However, a speech rendering is temporal and requires the user to issue commands to direct the browser to read and reread content. The user could issue a command to speak the entire content of a Web page, but in general this is an inefficient way for the user to locate the information they are interested in. It is the equivalent of a sighted user reading the entire contents of a Web page before they started looking for links or other groupings of information on the page. Most able-bodied users scan the screen for highlighted text and other visual cues to identify links and the grouping of information in the Web page. In a well-designed Web page the author has intentionally created cues to help users focus their attention on information the author thinks is

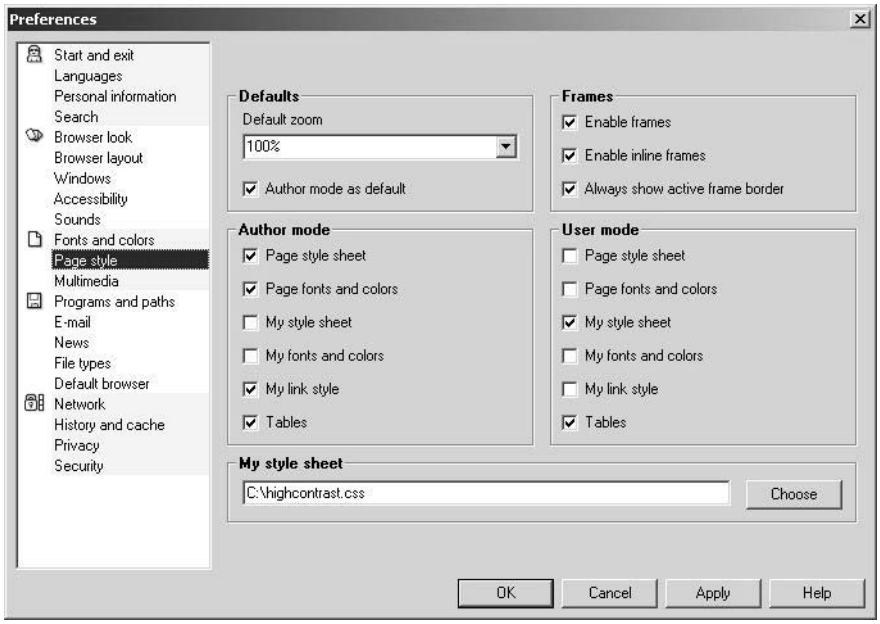


Figure 5: Opera 6.0 settings for rendering style preferences used in author and user modes.

important. The same is true in speech browsing. If authors include structural markup, users using any technology can style that structure to highlight the information to the user by way of speech, text, or graphical renderings.

Speech browsers like IBM Home Page Reader (http://www.ibm.com/able) and Freedom Scientific's JAWS screen reader (http://www.freedomscientific.com) have features for users to navigate HTML structural information. For example, they can navigate to elements marked as headings, links, and form controls, and through table data cells. These functions only work when authors use HTML header and other markup correctly in their Web resources. Figure 6 shows the read menu in IBM Home Page Reader (HPR), which highlights the different ways that HPR can be used to navigate through content. By providing users access to the structured markup it is easier for users to find the main groups of content without reading the entire document. Speech

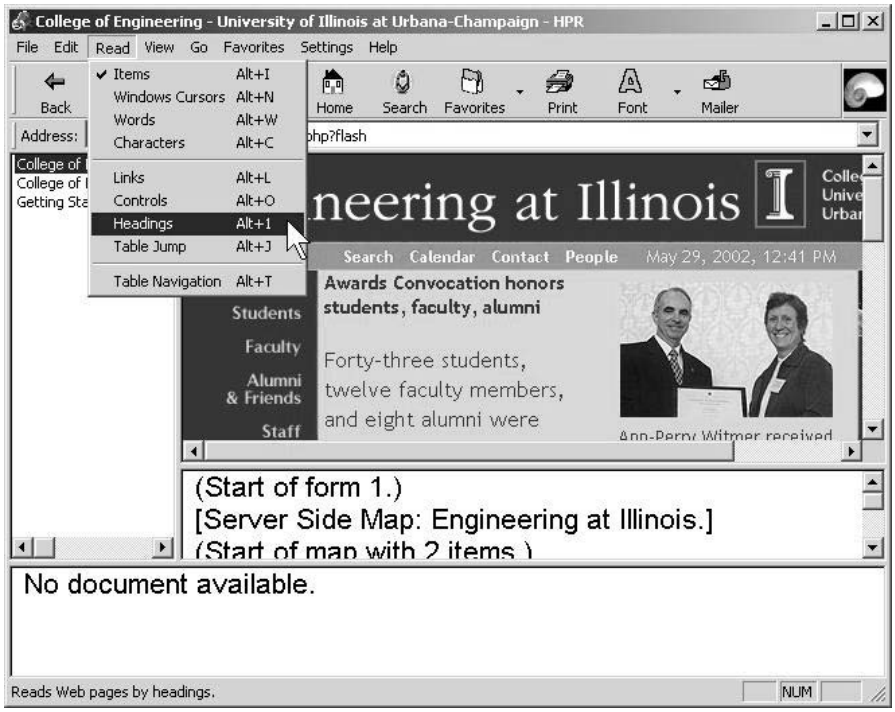


Figure 6: Main menu reading options in IBM Home Page Reader.

can be styled to indicate different types of elements. In IBM Home Page Reader the reading voice for links is styled as a female voice while non-link content is styled as a male voice for static text (voice can be configured to other settings by the user). Form controls that have explicit text labels associated with them can have their labels announced when the form control receives focus.

Many Web pages do not contain structural markup. This forces speech browsers to look for implied structure if they want to offer the user more than just a linear reading of the document. In the example of form controls, speech browsers may try to calculate the relationships between the form control and text around the form control to determine the label for the control. This is problematic since the guess can be wrong and instead of helping the user, the user maybe confused over the purpose of the control. If images do not have ALT text descriptions, current speech browsers may use the file name of the image in hopes that it may contain some useful information about the purpose and the content of the image. These approaches may help accessibility when the calculations are correct. When calculations are not correct they can seriously hurt accessibility by increasing the confusion and misinformation to users, which results in users taking more time to explore the resource or improperly completing the task they are trying to complete. In general these types of calculations should be unnecessary if authors and authoring tools (Treviranus et al., 2000) supported the inclusion of information for accessibility.

WEB DESIGN GUIDELINES

The W3C Web Content Accessibility Guidelines 1.0 (Chisholm, Vanderheiden, & Jacobs, 1999) and the U.S. Government Section 508 regulations for Electronic and Information Technology Accessibility Standards (Access Board, 2000) are the most widely used Web design requirements for designing or repairing Web resources to be more accessible. The W3C Guidelines were designed from the perspective of the needs of people with disabilities and was a public and consensus-based process. The Section 508 Web requirements were based on a need of the U.S. Federal Government to define Web accessibility standards that they felt were achievable and verifiable given the current state of disability access to the Web and the limitations of industry understanding of how to support the accessibility of web resources by people with disabilities. The Section 508 requirements were loosely based on the Priority 1 requirements of the W3C Guidelines, so the requirements are similar at that level.

W3C WEB CONTENT ACCESSIBILITY GUIDELINES

A detailed review of the W3C requirements (Chisholm et al., 1999) is beyond the scope of this chapter, since there are 64 checkpoints (requirements) organized into 14 guidelines with each requirement having an associated priority. The priorities associated with each of the WCAG checkpoints correspond to the relative importance of satisfying the requirement to the needs of people with disabilities. This section will highlight the organization of the guidelines and the major accessibility themes.

WCAG Priorities and Conformance

The following are the definitions of the priorities used in WCAG to identify how important a particular requirement is for people with disabilities to access content:

Priority 1: One or more groups will find it *impossible* to access information in the document if this requirement is not satisfied.

Priority 2: One or more groups will find it *difficult* to access information in the document if this requirement is not satisfied.

Priority 3: One or more groups will find it *somewhat difficult* to access information in the document if this requirement is not satisfied.

An author of a Web document can claim conformance to the guidelines at one of the priority levels when they satisfy all the requirements at the desired priority level and the preceding levels. An author can claim Single A conformance to WCAG when all applicable Priority 1 requirements for their document are satisfied, Double A conformance when all Priority 1 and 2 requirements are satisfied, and Triple-A conformance when all Priority 1, 2, and 3 requirements are satisfied. The publication of a conformance claim is useful in promoting accessible design and providing users with expectations on the accessibility of the resource.

WCAG Guidelines

The guidelines are organized into logical groupings of requirements. The document was developed on the notion of HTML and CSS technologies, but the accessibility requirements were intentionally designed to be more generic to allow the requirements to be applied to other W3C and non-W3C technologies as they became available. The WCAG requirements do not have the specificity of requirements for HTML defined in the Section 508 Web accessibility guidelines. The W3C guidelines do have an associated techniques document (Chisholm et al., 1999) that does provide design examples for HTML, CSS, and other technologies. The techniques document is designed to be informative and authors are not required to use the techniques outlined in the techniques document to satisfy a particular WCAG requirement.

Guideline 1: Provide Equivalent Alternatives to Auditory and Visual Content

Images need to have text descriptions for people with visual impairments and some types of visual-processing learning disabilities to understand the purpose and the content the image conveys in the document. Some images may require longer descriptions to convey the same information to the user. For example, images of a famous painting or a satellite photograph would need a longer description if the author uses the image to convey important information in the document. An image of a numerical chart or table would need to have a text table of the same

information available. Audio resources with speech need text transcriptions and videos need to be captioned and have text descriptions of action.

Guideline 2: Do Not Rely on Color Alone

Some Web designers encode information in color. There are many people with visual impairments (including people with color blindness) that cannot see the information coded as a color. In this case there needs to be another means to convey the information. Examples of information based on color are if the color of a score on a test was used to indicate the letter grade associated with the score or there were directions on the Web page to press a red button. To correct this problem the actual letter grade should be included with the scores and the ALT text label of the button should be referred to instead of the color of the button.

Guideline 3: Use Markup and Style Sheets and Do So Properly

One of the main problems with current Web design is that authors use graphical styling to encode the structure of the document. Instead of using the HTML H1 element to indicate the main topic of a document, authors use the FONT element to style the text for the main topic or an image of prestyled text. So users who cannot use the author styling and apply their own styling or are using a speech rendering will not be able to identify the main topics of the document. When using HTML the proper way to indicate structure is to use the HTML elements like H1–H6, LABEL, CAPTION, TH, and MAP and the list elements (OL, UL, DL) to properly indicate the structure between elements of information in the document, and to use Cascading Style Sheets to style the elements for different types of visual effects.

Validating HTML markup is also important in making sure that documents meet the requirements of the HTML language, which ensures that documents can be rendered on any HTML compatible browser. Many times authors or authoring tools use proprietary features of a particular browser or create invalid markup that can only be rendered when using the HTML repair features of one or two browsers. This limits the choices users have for accessing the content. Many times authors are not even aware that they have created content that can only be rendered in one or two browsers. Creating valid HTML markup will become more important as the Web matures and XML becomes more widely supported. Browser developers will want to focus their energies on exploiting the capabilities of XML and not repair invalid HTML markup. HTML is being replaced with then newer XHTML, which provides a consistent markup that is inherently more accessible. Slowly these repair features will disappear from browsers, like in the move from Netscape 4.7 to Netscape 6.2+ which is designed to support W3C standards.

Guideline 4: Clarify Natural Language Usage

This requirement is critical for speech browsers, since the only way a speech browser can identify the language reliably is when the author adds language information to the document. In HTML every element can include a LANG attribute to indicate the language of the content of the element. For example a document that is primary English should use <HTML LANG = "en"> in the beginning of the document. If the author uses a French quotation, the container element for the quote should include a LANG = "fr".

Guideline 5: Create Tables that Transform Gracefully

Most of the HTML table markup used on the Web is for graphical positioning. This is a potential problem for speech renderings that read information in document order. Table formatting that puts connected information out of document order can be confusing to speech users or people who use technologies that do not render table markup (i.e., Lynx browser). If tables are used for layout they should be as simple as possible and should be tested with a speech browser, a text-only browser, or a graphical browser like Opera that can be configured to ignore table markup to verify that the linear rendering makes sense when the table markup is removed.

Guideline 6: Ensure that Pages Featuring New Technologies Transform Gracefully

Technologies like Macromedia Flash, Adobe Acrobat, and XML technologies like MathML, SVG, and WAP have varying degrees of disability access solutions, so when using technologies like these you will need to determine the current extent to which the technology supports users with disabilities. Many times technologies will not be able to meet the needs of major disability groups and alternatives that provide a more accessible version of the information will need to be created. This type of information redundancy should not be considered a problem necessarily, but as an opportunity to provide information in more than one form that provides everyone with the opportunity to use information that is in a form most useful to them and their needs, which is the original pupose of the Web.

Guideline 7: Ensure User Control of Time-Sensitive Content Changes

People who need extra time to read information or who have physical impairments that slow their response time need to be able to have additional control over time-sensitive information. Providing mechanisms for the user to receive extra time to respond to a prompt is important and should be an option on pages with time-sensitive input. In secure environments it would be useful to allow user settings or configuration options to provide the extended response information throughout the system.

Guideline 8: Ensure Direct Accessibility of Embedded User Interfaces

Embedded technologies like Java and Active-X need to be compatible with assistive technologies, but also have built-in accessibility features. This may require adding additional controls to allow the user to style text and other objects presented through the embedded interface; or to provide an option for the user to style the interface based on the user's operating system style preferences. Keyboard support is important in the design of embedded user interfaces and the user needs to be able to control automated behaviors. Many times technologies will not be

able to meet the needs of major disability groups and an alternative that provides a more accessible means for people to access the information available through the embedded interface will need to be created.

Guideline 9: Design for Device Independence

One of the major problems for dynamic Web content is that designers only include support for pointer devices. It is important to use device-independent events or redundent event handlers to allow users to interact with the content using the widest number of input devices possible, including only the use of the keyboard. When device-independent event handlers are not available, make sure that you at least support the keyboard and mouse pointer for all the functionalities of your dynamic content.

Guideline 10: Use Interim Solutions

There are gaps between what browser and assistive current technologies can offer for accessibility and what specifications provide as accessibility features. This is by nature a dynamic requirement, so the requirements in this section should fade as technologies become obsolete.

Guideline 11: Use W3C Technologies and Guidelines

The use of W3C technologies are recommended since recent W3C specifications have been reviewed for accessibility features and support open and interoperable standards. This means that people using W3C technologies can support users with disabilities and also provide users with more choices to access content. Technologies like Adobe Acrobat and Macromedia Flash are adding accessibility features, but their features are based on retrofitting their current technologies with accessibility features. The retrofitting process often limits the capabilities of their players for rendering information accessibly, because the technology may have inherit accessibility problems due to the original design of the technology. In contrast technologies like HTML, CSS, and SMIL are supported by many developers and give the user more choices in accessing content.

Guideline 12: Provide Context and Orientation Information

One of the primary problems in current Web-site design is the lack of information that can be used by nongraphical renderings to identify the structure and the relationships of information on the page. Many Web authors view the Web as primarily a graphical medium and use graphical methods to encode the structure of the document. These graphical techniques do not translate the structural relationships very well to text and speech renderings. Often the graphical techniques used to indicate structure actually cause information to be distorted in nongraphical renderings, which typically use document order as the means to render information. If table markup is used to position information for a graphical rendering the document order often separates connected pieces of information, making the nongraphical rendering confusing.

Guideline 13: Provide Clear Navigation Mechanisms

Navigation is an important issue, especially for accessing information in Web sites, documents that have a large number of links, or large structured documents. Some examples of how markup can be used to improve navigation in HTML include these:

The text associated with a link to indicate the destination of the link,

The use of markup to provide users with a means to skip over repetitive navigation links,

Use of the MAP element to indicate a collection of related links, and

Use of the LABEL element to indicate the purpose of a form control.

Guideline 14: Ensure that Documents Are Clear and Simple

A requirement to use clear and simple language and layout is often very subjective and is often linked more for usability than for disability access. However, since many people with cognitive disabilities may have language impairments it is important to carefully review the terms and organization of Web resources to make the resources as easy to read as possible. Carefully consider the types of people who will be using the Web resources and their tasks and interests.

It is important to look at Web resources from the perspective of users and not from managers and other employees. One of the largest problems in Web site design is that many people design to meet their own needs, or the desires of the sponsors of the Web site or the manager of the organization the Web site represents. This often results in designs that do not meet the very different needs of the intended users. People within the organization usually understand procedures and relationships that users coming to the Web site do not. This often results in too much information on the main page, jargon unfamiliar to the users, and the expenditure of resources on visual effects that increase Web site visual esthetics, but do little to help the user to understand and complete tasks on the Web resource.

WCAG 2.0 Development

WCAG 2.0 is currently under development and this will supercede the current WCAG 1.0 requirements. For more information on the current status of WCAG 2.0 or to participate in the group activities go to their home page: http://www.w3.org/WAI/GL.

U.S. SECTION 508 REQUIREMENTS

The Section 508 Web Electronic and Information Technology Accessibility Standards (Access Board, 2000), developed by the Access Board of the U.S. Federal Government, includes accessibility requirements for all electronic machinery, computers, and software used by the federal government. Other regulations of the Access Board have been interpreted and applied to information in the Americans with Disabilities Act (1990) technical requirements.

The Web accessibility requirements of Section 508 are based mostly on the W3C Web Content Guideline Priority 1 requirements and a few additional requirements defined

by the access board. Therefore Section 508 requirements are considered a minimum accessibility requirement, basically ensuring that a Web resource is not impossible for a person with a disability to access. One of the organizational differences between WCAG and Section 508 is that the Section 508 requirements are designed to be much more specific to HTML and CSS technologies. Section 508 also has an additional requirement on functional performance, which is a general requirement for all those of Section 508, but applies also to the Web requirements. The Web requirements for accessibility are minimal, and authors are encouraged to consider design features more accessible than 508 requires.

The following are the Section 508 requirements for Web content with comments related to specific WCAG 1.0 checkpoint requirements:

A text equivalent for every nontext element shall be provided (e.g., via "alt," "longdesc," or in element content) *(compatible with WCAG Checkpoint 1.1)*.

Equivalent alternatives for any multimedia presentation shall be synchronized with the presentation (*compatible with WCAG Checkpoint 1.4*).

Web pages shall be designed so that all information conveyed with color is also available without color, for example, from context or markup (*compatible with WCAG Checkpoint 2.1*).

Documents shall be organized so they are readable without requiring an associated style sheet (*compatible with WCAG Checkpoint 6.1*).

Redundant text links shall be provided for each active region of a server-side image map (*compatible with WCAG Checkpoint 1.2*).

Client-side image maps shall be provided instead of server-side image maps except where the regions cannot be defined with an available geometric shape (*compatible with WCAG Checkpoint 9.1*).

Row and column headers shall be identified for data tables (*compatible with WCAG Checkpoint 5.1*).

Markup shall be used to associate data cells and header cells for data tables that have two or more logical levels of row or column headers (*compatible with WCAG Checkpoint 5.2*).

Frames shall be titled with text that facilitates frame identification and navigation (*compatible with WCAG Checkpoint 12.1*).

Pages shall be designed to avoid causing the screen to flicker with a frequency greater than 2 Hz and lower than 55 Hz (*compatible with WCAG Checkpoint 7.1*).

A text-only page, with equivalent information or functionality, shall be provided to make a Web site comply with the provisions of this part, when compliance cannot be accomplished in any other way. The content of the text-only page shall be updated whenever the primary page changes (*compatible with WCAG Checkpoint 11.4*).

When pages utilize scripting languages to display content, or to create interface elements, the information provided by the script shall be identified with functional text that can be read by assistive technology (*no WCAG 1.0 equivalent*).

When a Web page requires that an applet, plug-in, or other application be present on the client system to interpret page content, the page must provide a link to a plug-in or applet that complies with §1194.21(a) through (l). *This is an important user functionality that is part of the W3C User Agent Accessibility Guidelines 1.0, instead of WCAG 1.0.*

When electronic forms are designed to be completed online, the form shall allow people using assistive technology to access the information, field elements, and functionality required for completion and submission of the form, including all directions and cues (*requires more than UAAG Checkpoint 10.2 and 12.4*).

A method shall be provided that permits users to skip repetitive navigation links (*no WCAG 1.0 equivalent*). *Skipping repetitive navigation links is considered an important usability feature to help users skip over repetitive navigation bars and advertisements to get to the main content of a document faster.*

When a timed response is required, the user shall be alerted and given sufficient time to indicate more time is required (*no WCAG 1.0 equivalent*). *This is an important user functionality, and this problem is addressed in the W3C User Agent Accessibility Guidelines 1.0, instead of WCAG 1.0 by the W3C.*

EVALUATION AND REPAIR TOOLS

The Section 508 requirements and W3C WCAG guidelines can be rather tedious to use and to many authors the terminology used in the guidelines is unfamiliar. Automated analysis and repair tools have been developed to help authors identify the accessibility problems in their Web sites.

HTML VALIDATION

HTML validation is not usually considered an accessibility check, but valid HTML documents help accessibility in two ways. First valid documents are more reliably rendered in a wider range of technologies, including specialized technologies for people with disabilities. People with severe disabilities often need to use less popular or custom technologies to access Web information. The second way is that HTML and XHTML have requirements that support accessibility. For example, the inclusion of an ALT attribute (short text description of the image) for IMG and AREA elements is required for a document to be valid, which is one of the most common accessibility problem on the Web. Figure 7 shows the HTML validator service of the W3C (http://validator.w3.org/).

EVALUATION TOOLS

The first generation of automated accessibility evaluation tools is exemplified by the Bobby Web site and software (http://bobby.watchfire.com). Bobby was originally developed by the Center for Applied Special Technology (CAST) and the technology was transferred to Watchfire (http://www.watchfire.com) in the summer of 2002. Bobby can provide an evaluation of a Web resource based on either the Section 508 or WCAG requirements. Bobby

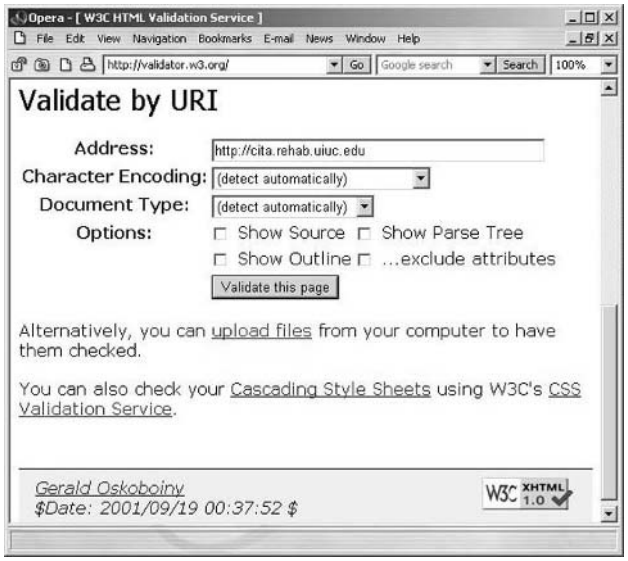


Figure 7: Image of the Web-based HTML validation service offered by the W3C.

analyzes the markup used by the author and reports on known or potential accessibility problems. Known problems are easily identified when markup is missing, for example, the ALT attribute on an IMG element. Other problems like the accessibility of scripting or the use of multiple languages in a Web document cannot be easily determined through a computational analysis and require the author to manually determine whether there is a problem. In this case Bobby indicates to the user a manual check is needed to determine accessibility. An example report from the Web-based version of Bobby can be viewed in Figure 8.

The advantage of using Bobby or a similar tool is that the author only needs to deal with the accessibility of the markup they are actually using. For example, if scripts are not used in a document the report does not include any information to check the accessibility of scripts. This helps the author to focus their attention on the problems of their particular Web-site design, essentially a custom set of guidelines for their design style. One of the limitations of tools like Bobby is when they check for the presence of markup, like ALT attributes for IMG elements, they cannot determine whether the ALT text content really represents the use or information the image conveys. For example, some authoring practices and automated authoring tools use the file name of the image as the ALT text to satisfy HTML validation requirements. While some tools will flag this as a potential problem, others just assume the ALT text is useful and do not report the problem to the evaluator. Tools like Bobby usually generate a large number of manual checks, which require the author to understand the accessibility requirement and do their own analysis of the markup to determine whether they satisfy the accessibility requirements. An example of a manual check is determining whether color is the only way some type of information is being represented (e.g., text labels for form controls in red to indicate a required response). Since some people may not be able to see those colors the information would not be accessible. The main advantage of the automation tool is that it narrows the scope of the manual checks to only the requirements

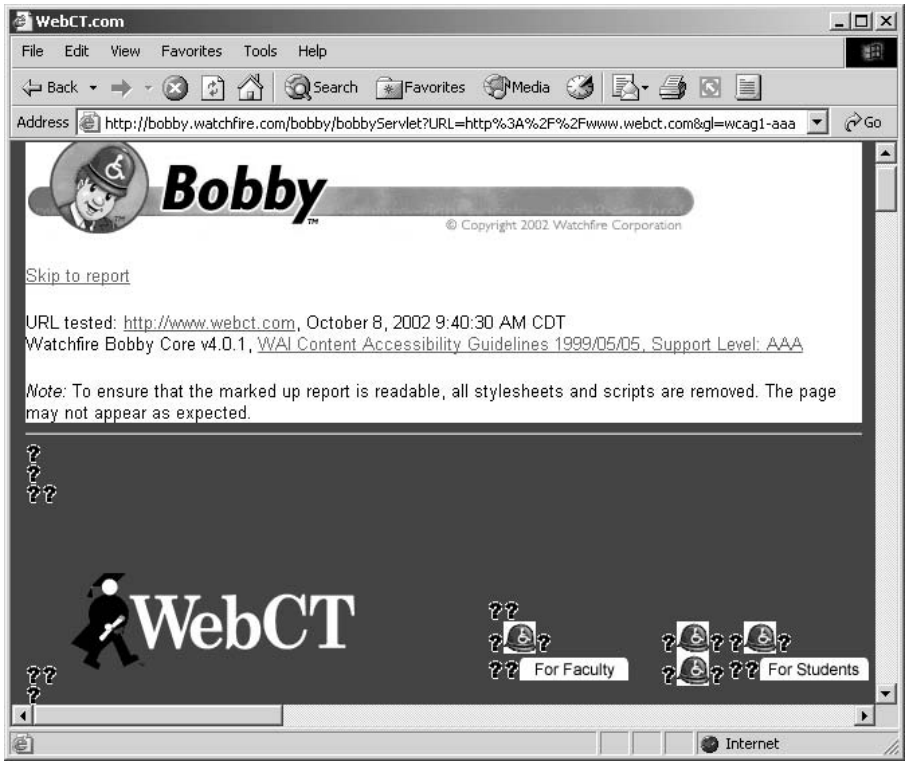


Figure 8: Image of an evaluation report generated by the Bobby Accessibility Evaluation Web resource.

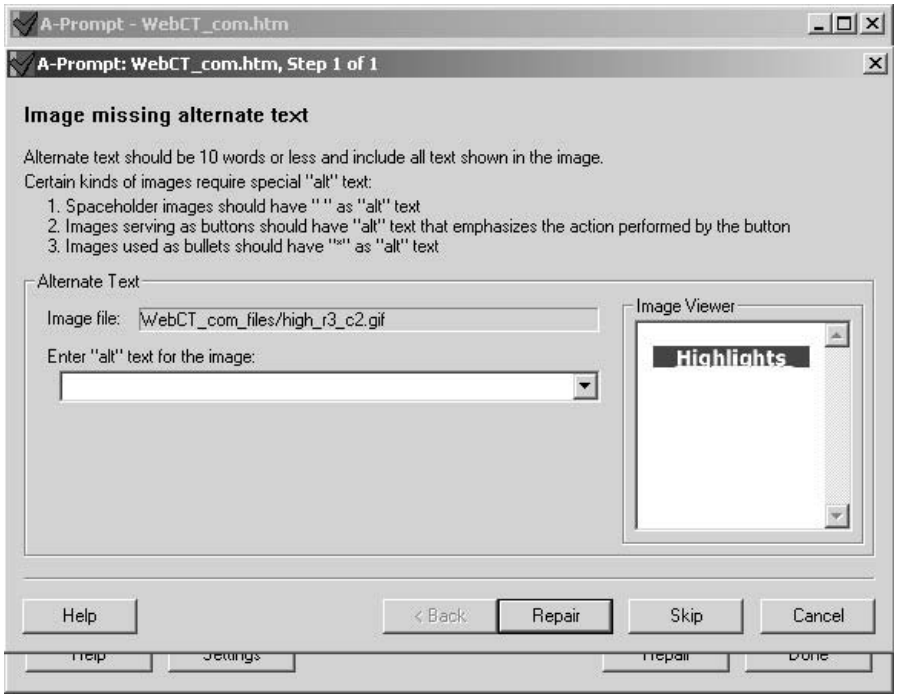


Figure 9: A-Prompt evaluation and repair tool.

the author needs to address based on the markup they used.

EVALUATION AND REPAIR

The second generation of accessibility evaluation tools provides both evaluation and repair services. In addition to identifying known or potential accessibility problems the second generation tools help the author repair the document. For example, if an image is missing the ALT text for an IMG element the repair tool can prompt the user to add the ALT text within the tool and add it automatically to the HTML markup. The ability to repair content directly in the tool improves the efficiency of repairing content, since the author does not need to go back into their original authoring tool to make the repairs. A-Prompt in Figure 9 is an example of a stand-alone tool that provides evaluation and repair services. Similar evaluation and repair tools have been developed for HTML authoring tools. LIFT from Usablenet (http://www.usablenet.com) is an example of a tool that works within Macromedia Dreamweaver and Microsoft FrontPage to help authors correct accessibility problems within the authoring tool. Figure 10 shows an example of the LIFT extension for Dreamweaver.

There are clear advantages to incorporating repair functions into accessibility automation tools, since authors receive additional guidance in repairing problems and they do not need to keep switching between the evaluation report and the authoring tool to make changes to the document. These types of tools assume that the user is developing a static Web page and has access to the source markup. Web resources generated through server side scripts and databases do not benefit from these types of tools, because the HTML markup is generated by the server each time a user makes a request to the URI. Also note that some HTML markup can be repaired through simple prompts (like missing ALT text), and other repairs will require more extensive revisions to the content than most repair tools can offer.

LIMITATIONS OF CURRENT AUTHORING TOOLS

The need for automated accessibility evaluation and repair tools indicates a severe weakness in current HTML authoring tools in helping authors intrinsically create accessible materials by default rather than by exception. Ideally authoring tools should make it easier for people to create universally accessible Web resources and guide them in the use of markup that supports accessibility. Currently authors need to have information on accessible design to create accessible content with most HTML editors. Some authoring tools actually impede the ability of the author to create accessible information. For example, in Dreamweaver MX (or earlier), the author cannot easily use the MAP element (typically, but not limited to use for image maps) to indicate a collection of related text links (i.e., a navigation bar). Dreamweaver warns the user that this is an invalid use of the markup (apparently Dreamweaver has been designed to only support AREA elements in a MAP container), and the text links can only be hand-edited into the code and are not rendered in the graphical preview of Dreamweaver. However, it is structural markup like MAP that is critical for making the next generation of Web technologies more accessible. Dreamweaver in this case actually makes it harder to move to the next level of accessibility.

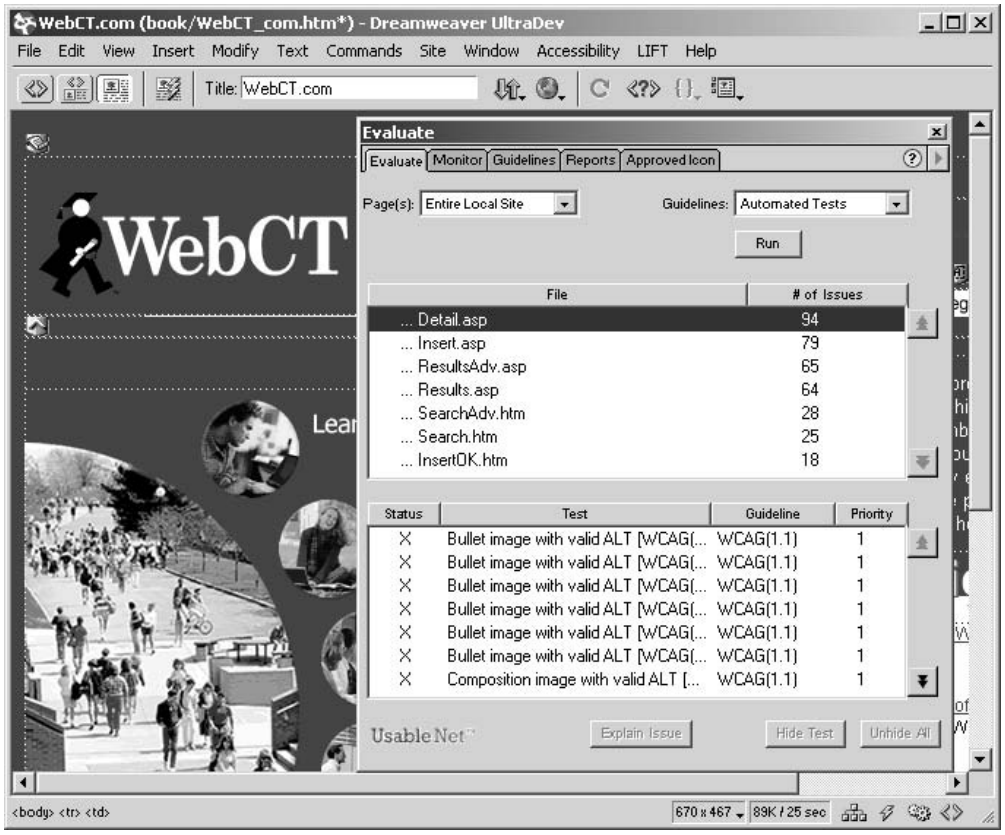


Figure 10: LIFT evaluation and repair tool integrated into Macromedia Dreamweaver.

MICROSOFT POWER POINT ACCESSIBILITY PLUG-IN

Not all authors use traditional HTML authoring tools to create content. For example, in educational institutions the most popular Web authoring tools used by instructors is Microsoft Office. Instructors can use the familiar features of MS Office products and save the content in an HTML-like format. Unfortunately the default Save features of MS Office produce XML content that can typically only be viewed in Internet Explorer. Authors who are more inquisitive can change the default settings to publish as HTML that can be viewed on other browsers, but this requires additional skill and knowledge on the part of the author. In either case the content developed is not accessible, and in this case the author does not even have an option to repair the markup, since Office tools do not provide a means to even hand edit the resulting HTML (or XML in some configurations of Office) code.

While the current situation with Office is not very good for publishing accessible Web content, there is tremendous potential with new types of tools to automatically generate accessible content from Office documents. By using the application programming capabilities of Office new save options can be added to the Office menus to create accessible content and can guide authors into adding additional accessibility information. An example is a tool that can convert Microsoft Power Point Slides into accessible HTML (http://www.rehab.uiuc.edu/office). It automatically creates parallel linked versions of HTML slides. One set of slides uses primarily text and CSS to provide a highly user customizable version of the slides and the other is the more traditional graphical version of the slides. Each version of the slide is linked to the other so the user can easily move between a graphical and a text view of the slides. This illustrates another important Web and accessibility concept: giving users the choice on how they want to view information. Unlike print materials which become more expensive or inconvenient to provide multiple views of the same information, there is little cost on the Web. Users can therefore pick the view that works for them based on their own needs and the task they are trying to complete. The tool also prompts authors for additional information when needed for accessibility. However, unlike current evaluation and repair tools the prompts are nontechnical and ask the user for the information, hiding the HTML coding details. Figure 11 shows the prompt for creating a text equivalent for an image. The user is asked about how the image is used in the presentation and then guides the author in creating a compatible text equivalent.

ACCESSIBLE REPAIR OR UNIVERSAL DESIGN?

The main characteristic of both Section 508 and W3C Web Content Accessibility Guidelines is that they base their requirements on a model of an author having already prepared the Web materials (an existing Web site, for example) and is trying to repair the Web site to be more

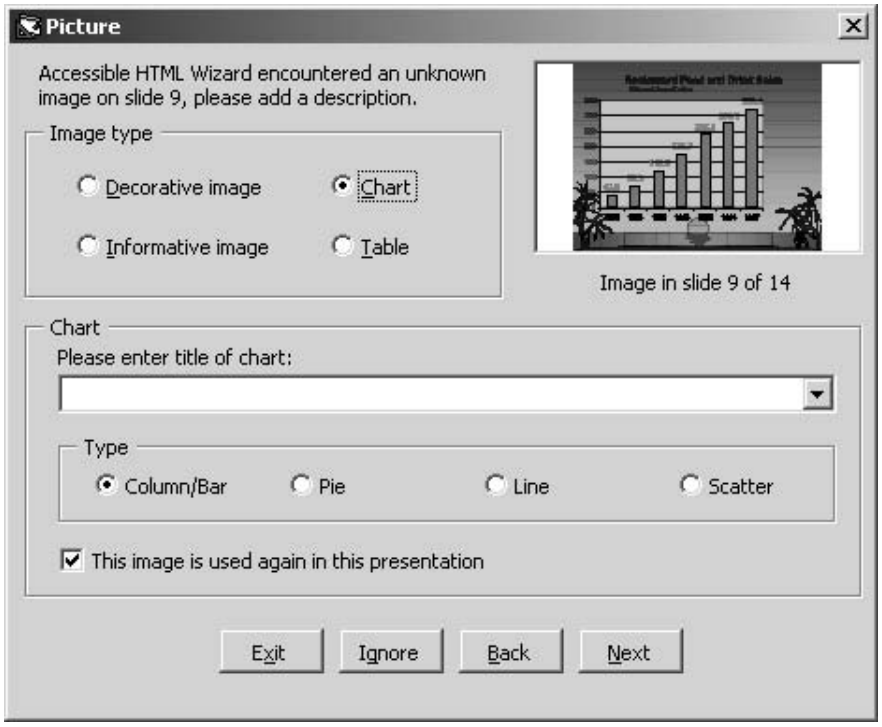


Figure 11: Power Point Web Accessibility Plug-in asking the user the purpose of an image in the presentation.

accessible. An alternative approach is based on universal design and making HTML design choices at the beginning of the design process or totally redesigning the Web resource with more accessible HTML markup techniques. The primary philosophical change for most authors to move to a universal design model of Web development is to move away from a graphical view and toward an information view where the author has preferred rendering styles, but makes it easy for the user to adapt the rendering to their technology, personal needs, and preferences. This includes mobile technologies like cell phones and PDAs using speech interaction techniques or character-oriented displays similar to the technologies used by people with disabilities.

A detailed discussion of universal design of Web resources is beyond the scope of this chapter. The following is a summary of the general approach for the development of static HTML documents (no scripting effects):

Do not use images to stylize text; instead use text with CSS styling (including creating visual button effects).

Use HTML header markup (H1–H6), correctly nested, to indicate new sections and their subsections.

Use HTML MAP element to indicate collections of related text links (navigation bars).

Use HTML list markup to indicate ordered or unordered lists of information and use CSS list styling to customize bullets and numbering.

Markup document language and language changes with the LANG attribute.

Use HTML LABEL element to indicate the text labels associated with form controls.

Use HTML TABLE markup sparingly for layout positioning elements on a page.

Use CSS background color and image capabilities instead of images for creating background effects.

Do not use images or CSS absolute position features for positioning; instead use CSS margin and padding to position information within simple layout tables.

Provide text equivalents for all nontext content (i.e., images, audio, and video).

Use the TH element and SCOPE and HEADER attributes to indicate header cells in data tables, and refer to them in the associated TD elements that contain the data.

Use only valid HTML and CSS techniques; do not support proprietary extensions of any particular browser and validate your documents before publishing them.

When these techniques are used it makes the resource not only accessible to people with disabilities, but provides all users with more flexibility to access content.

LAWS AND REGULATIONS

Many countries including the United States have started legally requiring Web accessibility for government and public Web sites (Thatcher et al., 2002). In the United States the Section 508 Web Accessibility requirements apply to federal Web resources. There is a weak provision in the rule that any state receiving technical assistance funds would also need to comply with the Section 508 requirements for state agencies. Since states were not involved in the design of the requirements, it is likely most states would return the money to remove the requirement if they

felt it was a burden. Under ADA and Section 504 the Rehabilitation Act of 1973 both call for "effective communication" of information in a timely manner. Since the Web is now a major means of communication of information it would be hard to argue that providing "effective communication" of Web information could be done effectiently through some other medium or technology, since the Web provides 24-hour by seven-day-a-week access to information. Therefore the main question is what standard should be used to determine whether a Web resource is accessible. Right now the Section 508 requirements would probably be considered the minimum in the United States since the federal government has adopted them. These requirements are often not enough to provide effective communication. For example, Section 508 does not have a requirement for including language information. Language markup is needed for Web sites with more than one language for speech output systems to know when to switch to speaking another language. Without this most multilanguage information Web sites are not accessible. These types of Web sites are commonly used in on-line language foreign language education courses. Additional litigation will probably need to occur before the legal requirements of accessibility are clearly understood.

CONCLUSION

The universal design of Web content not only provides users with disabilities access to Web content, but all users will have more choices and more control over the rendering of it. Just like concrete curb cuts and ramps have benefited the general population in many ways, the electronic Web accessibility curb cuts and ramps will benefit all users of the Web.

GLOSSARY

ADA U.S. American with Disabilities Act of 1990 that guarantees the rights of people with disabilities to have access to public spaces, services, and employment.

ATAG W3C Authoring Tool Accessibility Guidelines provide authoring tool developers with information on how to support the creation of accessible Web content and be more accessible to people with disabilities.

ALT text The short description associated with an image on the Web, in the form alt = "brief description of the image."

CSS (Cascading Style Sheet) Cascading Style Sheets is a W3C technology designed to style HTML and XML content for rendering in a browser. The specifications began with CSS1 in December 1996, evolved to CSS2 in May 1998, and is currently under development as CSS3.

Disability A visual, hearing, muscular, learning, or mental impairment that substantially limits one or more of the major life activities of an individual.

HTML structure Using the structural markup capabilities of HTML to indicate the relationships between information in a Web resource.

Keyboard shortcuts The ability to use keyboard commands to control software, providing an alternative to pointing with a mouse to select functions.

LONGDESC A URI to a detailed description of an image and is an attribute of the IMG element.

Section 508 U.S. Section 508 rules and regulations are designed for use by federal agencies to provide access to services by citizens and accommodations to federal employees with disabilities. In December 2000 the Electronic and Information Technology Accessibility Standards included Web accessibility requirements.

Screen magnifier A software program that magnifies text and graphics, and controls the colors on a graphical computer system.

Screen reader A software program used by people who are blind to have the elements on a computer screen read to them through synthesized speech or refreshable Braille display.

Text equivalent A text description associated with nontext content like images, audio, and video.

UAAG W3C User Agent Accessibility Guidelines 1.0 provide information to developers of browsers and multimedia players on how to make their technologies more accessible.

Universal design The design of resources to adapt to the needs and capabilities of a wide range of users, including people with disabilities.

WAI W3C Web Accessibility Initiative is a program of the W3C to promote the accessibility of the Web to people with disabilities through education, design guidelines, and review of Web technologies for accessibility features.

WCAG W3C Web Content Accessibility Guidelines 1.0 provide information to Web content authors on how to create accessible Web materials.

CROSS REFERENCES

See *Cascading Style Sheets (CSS); Digital Divide; Electronic Commerce and Electronic Business; HTML/XHTML (HyperText Markup Language/Extensible HyperText Markup Language); Human Factors and Ergonomics; Legal, Social and Ethical Issues; Web Site Design.*

REFERENCES

Access Board (2000). *Electronic and Information Technology Accessibility Standards*. U.S. Architectural and Transportation Barriers Compliance Board. Federal Register, 36 CFR Part 1194, December 21, 2000. Retrieved June 6, 2002 from Access Board Web site: http://www.access-board.gov/sec508/508standards.htm

Americans with Disabilities Act (1990). *U.S. Public Law 101–336: The Americans with Disabilities Act*. Retrieved June 6, 2002 from U.S. Department of Justice Web site: http://www.usdoj.gov/crt/ada/pubs/ada.txt

CITA Surveys (2001a). *Web masters brown bag survey result*. Retrieved June 6, 2002 from University of Illinois at Urbana–Champaign, Illinois Center for Instructional Technology Accessibility Web site: http://cita.rehab.uiuc.edu/survey/web-masters-survey-result.html

CITA Surveys (2001b). *ADA Web accessibility workshop survey result*. Retrieved June 6, 2002 from University of Illinois at Urbana–Champaign, Illinois Center for

Instructional Technology Accessibility Web site: http://cita.rehab.uiuc.edu/survey/ADA-survey-result.html

Chisholm, W., Vanderheiden, G., & Jacobs, I. (1999). *W3C Web content accessibility guidelines*. Retrieved June 6, 2002, from the World Wide Web Consortium (W3C) Web site: http://www.w3.org/TR/WCAG10/

Chisholm, W., Vanderheiden, G., & Jacobs, I. (1999). *W3C Web content techniques document*. Retrieved June 6, 2002, from the World Wide Web Consortium (W3C) Web site: http://www.w3.org/TR/WCAG10-TECHS/

Jacobs, I., Gunderson, J., & Hansen, E. (2002). *W3C user agent accessibility guidelines*. Retrieved June 6, 2002, from the World Wide Web Consortium (W3C) Web site: http://www.w3.org/TR/UAAG10

Kay, H. S. (2000, July). Disability and the digital divide. *Disability Statistics Abstract, 22*.

Lie, W. L., & Bos, B. (1997). *Cascading style sheets*. Reading, MA: Addison-Wesley.

Pernice Coyne, K., & Nielsen, J. (2001). *Beyond ALT text: Making the Web easy to use for users with disabilities*. Retrieved June 6, 2002, from the Nielsen Norman Group Web site: http://www.NNgroup.com/reports/accessibility

Thatcher, J., Bohman, P., Burks, M., Henry, S. L., Regan, B., Swierenga, S., Urban, M., & Waddell, C. D. (2002). *Accessible Web sites* (Chapter 2). United Kingdom: Glasshaus Ltd.

Treviranus, J., McCathieNevile, C., Jacobs, I., & Richards, R. (2000). *W3C authoring tool accessibility guidelines*. Retrieved June 6, 2002, from the World Wide Web Consortium (W3C) Web site: http://www.w3.org/TR/ATAG10

U.S. Census Bureau (2001). *American with disabilities: Household economic studies*. Washington, DC: U.S. Department of Commerce.

FURTHER READING

Illinois Center for Accessible Instructional Design: http://cita.rehab.uiuc.edu

Trace Center: http://www.trace.wisc.edu

W3C Web Accessibility Initiative: http://www.w3.org/WAI

Access IT: http://www.washington.edu/accessit

Information Technology Technical Assistance and Training Center (ITTATC): http://www.ittatc.org/

National Center on Accessible Media:http://ncam.wgbh.org/

CAST: http://www.cast.org

Webable: http://www.webable.com

Usablenet: http://www.usablenet.com

Adobe: http://access.adobe.com

IBM: http://www-3.ibm/able

Microsoft: http://www.microsoft.com/enable

SUN: http://www.sun.com/access/

Unix Operating System

Mark Shacklette, *The University of Chicago*

INTRODUCTION

Unix is a powerful operating system that began life in the late 1960s and has continued to exert a powerful influence on operating systems to this day. The Unix operating system is known as a *multiuser, multitasking* operating system. This means that more than one user can be logged into and execute multiple programs on the system at the same time. Most people intuitively believe that personal computers were first introduced by IBM in the late 1970s with the delivery of the IBM Personal Computer. In reality, computers were, for the most part, personal computers from the very beginning, in the sense that most computers only allowed one user to work on the computer at a given time. In addition, that single user would only be able to run a single program at a given time. This notion of one user running one program after another is a concept known as *batch sequential processing*. Programs that individual users would want to run would be scheduled as a "batch." The various programs in a *batch* would be run one after the other, until the batch was completed.

Before Unix, programs would be entered on punch cards, or "IBM cards," and these cards would be placed into a machine called a card reader. The system would then execute the program. Output from these programs, known as "jobs," would be in the form of paper printouts from a printer. A programmer would know the results of her program only when the job was completed and output had been produced from the printer. If a mistake (often called a "bug") was discovered, the bug would have to be resolved, new cards created, and the entire process executed again—a time-consuming, tedious, and laborious process by any standard.

Unix is an operating system written by programmers for programmers. Programmers were tired of having to work with punch cards, and having to wait, sometimes for hours or even days, for the results of their program. It was a waste of valuable time. How much nicer it would be to be able to enter programs directly into the computer by typing at a terminal and submitting the program directly to the operating system to run—to run immediately and to allow multiple programmers to do this at the same time. Enter Unix.

HISTORY

Early History (1969–1972)

The history of the Unix operating system is itself a story full of intrigue, lawsuits, corporate and federal politics, budgetary concerns, computer games, antitrust law, lawsuits, resourcefulness, and, when the going got rough, subterfuge. It is a story about a few very smart people armed with a few very good ideas. The story reads at times like any great spy novel, full of plots and subplots, intriguing personalities, battles for territory, winners and losers. It even has a chapter on the illegal and surreptitious underground transmission of contraband books. It is a story whose participants seem compelled, like some ancient bearded mariner, to tell and retell, if for no other reason than to remind ourselves and the world that good ideas are always possible, always to be valued, and sometimes, just sometimes, may win through pure unabashed brilliance. In the end, the story of Unix is an unlikely story, a story that at any of a number of points in its development might not have happened at all, were it not for the commingling of those few very good ideas in the minds of those few smart people. That, of course, and the subterfuge.

In the Beginning: Multics

"Plan to throw one away; you will, anyhow." These words are from a chapter in Frederick Brooks's seminal work on software development, *The Mythical Man-Month* (Brooks, 1995). His point is that the first attempt at something new in software is always a learning experience fraught with the unknown, and wherever "a new system concept or new technology is used, one has to build a system to throw away, for even the best planning is not so omniscient as to get it right the first time" (p. 116). In the case of the Unix operating system, the one that was thrown away (albeit reluctantly) was Multics, an acronym that officially stood for "Multiplexed Information and Computing Service" but which also stood for, according to one tongue-in-cheek oral tradition, "Many Unnecessarily Large Tables in Core Simultaneously." Multics (*not* capitalized) began life in 1964 as a joint collaboration between three institutions: General Electric, the Massachusetts Institute of Technology (MIT), and AT&T Bell Telephone Laboratories (BTL). It ended its life only in 2000, when the last Multics computer was turned off at the Canadian Department of National Defence.

Multics was an attempt to experiment with a number of new ideas in operating systems research. Before Multics, operating systems tended to be monolithic, written in a native low-level assembly language, with "flat" file systems that did not allow multiple users access. Many of these early operating systems could not do several things at the same time, and often resorted to running different user programs back to back, or sequentially, in a mode that is called "batch processing." One of the primary goals of the Multics operating system was to provide a time-sharing system that allowed multiple users to work on the system simultaneously with safety, security, and integrity. The original goal was to support several hundred simultaneous users (Daley & Dennis, 1968).

The Multics operating system was notable for attempting to provide among other things, within a single operating system, a shared-memory multiprocessing environment. But the three most interesting aspects of Multics for our purpose were its emphasis on using a high level language (PL/1) for its implementation (as opposed to a machine-dependent assembly language), its notion of extracting the command processor user interface (shell) from the kernel into a separate user (selectable) program, and its implementation of the first hierarchical file system. A hierarchical file system is a file system that supports a compositional structure, that is, directories are files that can contain other files, including other directories, a commonplace in modern operating systems today (thanks to Multics and later Unix) but a somewhat radical concept in 1964.

The result of this design was that only a small portion of the operating system had to be written in assembly language. Writing the operating system in a high-level language afforded the possibility of more easily porting the operating system over onto disparate hardware platforms. This notion of writing the operating system in a high-level language was a new one. This notion of portability was one of many features of Unix that was derived from Multics, others included the concept of making the command interface (also known as the shell) just "another user program like the others." Literal carryover in command naming, such as "ls" (list directory), "pwd" (print working directory), and the notion of an online help system ("man" in Unix) all derived from Multics.

The joint collaboration between General Electric, MIT, and Bell Labs was, in the case of the latter, centered at the Murray Hill, New Jersey, Computer Research Group's (CRG) laboratory. The Multics development machine at the labs was a dual processor General Electric 645 computer, a slow machine by today's standards, operating only a little faster than the original IBM PC. A 1000 Mhz Pentium III is roughly *2,000 times faster* than the GE 645.

All was not well with Multics and the joint venture. Before tuning, the system could only support a few simultaneous users (remember the goal of hundreds of simultaneous users). After tuning, by 1969, a Multics system could support about 30 users per processor. Nevertheless, performance was erratic and ran on expensive hardware; as a result, Bell Telephone Laboratories (BTL) had seen enough of the Multics effort and pulled out of the triumvirate in the spring of 1969. This greatly saddened a number people in the CRG, including one hyper-productive bearded individual by the name of Ken Thompson. After BTL pulled out of the Multics project, the GE 645 was packed up and carted off, and Ken and a couple of his colleagues, Dennis Ritchie and Rudd Canaday, were left without a toy.

This was the summer of 1969, remember, and the nation and the world was captured by the *Apollo* missions and just about to witness a frightening several days when Jim Lovell and crew aboard the *Apollo 13* flight were going to use the Moon's gravity to whip them back home to Earth. Ken Thompson had been playing around with a computer game he had written on the now-missing Multics machine. "Space Travel" was a space orbit simulator that cost Ken's department $50 to $75 an hour to play on the GE 645 timesharing system. Even with that expense, the Multics operating system running on the old machine could hardly keep up with the hyperbolic calculations that were needed to run the game efficiently. So Ken found himself simultaneously without an operating system to work on and without a computer on which to play Space Travel. This situation was, of course, entirely unacceptable. Something had to be done, and done fast.

A "Little-Used PDP-7"

After pulling out of the Multics project by BTL, Ken's group naively petitioned the powers that were at BTL for the purchase of a brand new DEC 10, at a cost of around $120,000. Why? they were asked. Their response was so they could write a brand-new operating system that looked a lot like Multics. The answer that came back somewhat bluntly basically said, "What part of *no* don't you understand?"

Ever resourceful and constitutionally incapable of fatigue, Ken found an old (and even outdated by that time) PDP 7 computer lying unused in the corner of an office and decided to port Space Travel onto it. The game ran, but very slowly. Space Travel was in desperate need of a file system to support its planet information, and with their experience of the Multics hierarchical file system, Dennis

Ritchie and Rudd Canaday had already begun sketching out a design for a file system of some theoretical new operating system that BTL was not going to allow them to develop. This file system design was compositional and had a root node from which descended either files or directories, which were just another type of file. Ken liked the design and implemented the file system over a single night on the PDP 7.

Having the PDP 7 provided a cost-free machine (someone else at BTL had paid the $72,000 for the computer) on which to develop that new operating system they wanted to work on. When Ken's wife decided to take their new baby boy out to California to visit Ken's parents for a month, Ken used that month to implement a new operating system kernel. In one month he produced a kernel, a shell, an editor, and an assembler to natively compile programs on the new kernel. In one incredible month, he had produced a new multiuser multiprocessing operating system. The moniker UNICS was given to the new operating system, reportedly by Peter Neumann (but some say Ken Thompson), as a pun on the word Multix: UNICS was punned the UNiplexed Information and Computing Service, a not-so-subtle implication that it was some form of "castrated" Multics. But it was hardly that.

There was a method to all this madness, however, because whereas Multics was a "serious" operating system, Ken, Dennis, and Rudd wanted to keep their new system *simple*. Simple meant manageable and comprehensible and extensible. They believed that one of the things that had led Multics astray was its complexity. The sheer smallness of the 18-bit PDP 7 machine forced an economy of vision, and certainly an economy of bits, and this economic vision produced the central philosophical tenet of Unix: *small is beautiful*.

The Sting

Even though the initial version of Unix was written in PDP 7 assembly language, it was still painfully slow. They had to have a new machine with the hardware resources capable of supporting the new multiuser and multiprocessing operating system they had developed and were envisioning. There was certainly a desire, again learned from Multics, to have the operating system written in a high-level language, for portability reasons. Current compilers such as Fortran were ruled out immediately because they were much too large. The GECOS machine (a GE 635) had a compiler that had been ported to it called Basic Combined Programming Language (BCPL). Even BCPL was too large for the PDP 7's fledgling new operating system. So the team cut out all but the essentials, and turned BCPL into an interpreter (which only generates intermediate code rather than machine code) and called it simply "B" to indicate that it was a highly abbreviated form of BCPL. It was a temporary solution, but not an ideal one, and, as it turned out, not a permanent one. What they really needed was a larger and more capable machine.

Knowing that the powers that were at BTL were going to be stricken with apoplexy on any proposal with the words "operating system" in it, Joe Ossanna, forever interested in text processing, reminded Ken Thompson that the Patent Department was in need of a text processing system, something akin to a program called "runoff" that had been deployed on another system. Ossanna suggested that the managers might actually accept a proposal that provided an actual solution to an immediate need, which was a text processor that could support specific terminals and provide line numbered output.

The Coup

A proposal was made and eventually accepted and a new computer, a Digital Equipment Corp. PDP 11/20, was provided for the development of the new text processing system for BTL's Patent Department. But an operating system was needed on which to develop the text processing system. What operating system did they choose to run on this brand new computer? Unix, of course! (The UNICS name soon morphed into the now familiar Unix form.) On November 3, 1971, K. Thompson and D. M. Ritchie published the first *Unix Programmer's Manual,* marking the first official release of the Unix operating system running on the PDP 11/20. The last sentence of the introduction to the manual simply read: "This manual was prepared using the UNIX text editor *ed* and the formatting program *roff*." The first released version of Unix supported three typists from the Patent Department at Bell Telephone Laboratories in Murray Hill, NJ. The roff text processing system was written literally over a single night. The rest of the time was spent on Unix.

The Patent Department liked the system and eventually bought the 11/20 machine from Computer Research Group. The funds received for the 11/20 were used toward the purchase of another faster and better machine: the DEC 11/45, onto which Unix was immediately ported.

Infancy and Politics

With the publication of the third edition of Unix in early 1973, the number of Unix installations had grown to 16, far beyond just the Patent Department. The fourth edition appeared later that year and provided a complete rewrite of the Unix operating system from assembly language into C, derived from B, a highly portable systems programming language that produced native machine code, developed primarily by Dennis Ritchie. This was a seminal event in the history of Unix, because for the first time, porting Unix onto a new hardware platform was significantly easier compared with trying to port an assembly-based operating system. Unix grew up within BTL like any inventive intrigue, by oral tradition and reputation for delivering the goods. And news of its abilities and philosophy spread. The fact that it actually *worked* didn't hurt either.

A Slight Marketing Problem

In October of 1973, Ken Thompson and Dennis Ritchie delivered a highly successful talk on the new Unix operating system to an Association of Computing Machinery conference on operating systems. The paper was titled "The UNIX Time-Sharing System" (Ritchie & Thompson, 1978). It gave a technical summary of the key components of the operating system. It also mentioned that the system was from the beginning self-maintaining through the fact that the source code was freely available to licensed users. This talk was met with enormous interest, and requests began pouring in for the opportunity to

purchase or otherwise obtain a copy of Unix. People wanted to buy Unix.

But AT&T couldn't sell it. Not for money, anyway. The reason for this goes back to 1949, and an antitrust complaint the Truman administration filed against American Telephone and Telegraph Company (AT&T) and the Western Electric Company. The net result of this was that AT&T could not do business in anything that didn't have to do directly with delivering phone service. Period. Additionally, AT&T had to provide its patents to the public and could only license (i.e., not sell) their technology to others at nominal fees. This basically meant "no business but phones and telegrams" (Salus, 1994, p. 58). Posing the question "But can't we still sell Unix?" to the lawyers would have been met with even more acrimony than the question "Can we develop another Multics?"

If You Build It, They Will Come

So AT&T decided to license Unix for a nominal charge to other interested parties. If others, like universities, want to play with this research, that's fine, they could do that. Give it to them, but don't market it, don't support it, and don't promise to fix any problems with it. By the end of the next year, the list of "other interested parties" read like a list of Academic Who's Who: Harvard, Columbia, Stanford, Princeton, UC Berkeley, and Johns Hopkins. Universities liked Unix: There was no warranty to void because there was no warranty; users got the source code, which made the system inherently fixable; it came with its own compiler, which allowed users to write whatever new utilities you wanted; it supported multiple users and multiple programs simultaneously; and it ran on hardware that was beginning to cost less than $20,000. Unix took off.

The Great Schism

In March of 1975, Digital Equipment Corporation introduced the powerful new PDP 11/70, and the Computer Science Department at the University of California, Berkeley, purchased one of the first ones produced. The following fall, Ken Thompson was invited to Berkeley as a visiting professor in the Computer Science Department. One of the first things he did was help Berkeley install Version 6 of AT&T Unix onto their new PDP 11/70. The other thing he did was meet a new PhD student by the name of Bill Joy.

Version 6 was a popular edition and was widely adopted by university computer science departments across the world. Operating systems classes began to be taught using the Unix operating system, and one of these classes was offered at the University of New South Wales, Australia, by a professor there named John Lions. He published a detailed commentary on the source code of Version 6, which helped countless programmers across the world repair and improve their installations of Version 6. Again, the widespread distribution of source code and now commentary further enhanced the interest in Unix and the practicability of its successful deployment.

Thompson had written the ed editor, which had been available since the first edition. It was a primitive line editor, certainly by today's WYSIWYG (What You See Is What You Get) word-processing standards. The ed editor was a command mode editor that allowed users to open and edit a file, make modifications, and save it back to the disk, all without ever seeing the text they were editing (this actually has significant practical benefits, because the ed editor is a filter and can be command-driven from a shell script, using what is known as a "here document").

Bill Joy didn't like ed very much. Most mortals prefer to be able to see the text they are modifying as they are making the changes. (Actually, ed did allow you to see the text, but only line by line and then only on request.) So Bill Joy wrote a new visual editor called "vi" (for "VIsual") as a front-end for ed. Vi also used some of the new terminals available and allowed the user to see a screen full of the text that he or she was editing. The source code for vi was sent to Ken and incorporated into the next AT&T Unix releases. Bill also didn't care for the default user shell and decided he would prefer a shell with a scripting language that more closely resembled native C syntax than did the existing shell, so he created the C Shell (csh).

The Berkeley Software Distribution

Bill Joy wasn't content with just writing a new shell and an editor. He soon turned his sights toward the V6 kernel itself. His initial focus was on improving the speed of the kernel. He released a "modified" Unix system in early 1978 through the Computer Systems Research Group (CSRG) at Berkeley and called it the Berkeley Software Distribution, or BSD 1.0 for short. BSD 2.0 appeared shortly thereafter, and included Bill's new vi editor along with his new C Shell. The first two releases of BSD added only user programs and shipped with the existing AT&T V6 kernel. 1BSD was sent out in return for a license payment of $50.

In early 1978, a group at BTL in Holmdel, NJ, had just received a Digital Equipment Corporation (DEC) VAX and had taken Version 7 and ported it over to the large VAX instruction set. About the same time, Berkeley received its own VAX, an 11/780, which was a big minicomputer that boasted more memory among other advantages. Berkeley contacted AT&T and requested that their Holmdel VAX port of Version 7 be made available to them, and it was. The Version 7 VAX port became the basis of 3BSD, which was made available in 1979, and shipped with a new virtual memory system (supported by the VAX; the "VA" stands for "Virtual Addressing") that the BSD group had implemented for the VAX, which allowed for the support of many new powerful applications.

BSD and the Internet

Unix systems prior to 1976 were pretty much standalone systems that couldn't talk to one another over a network. To address this, Mike Lesk of AT&T created a program that would allow a file to be copied over a modem from one Unix computer to another. This program was called uucp—Unix to Unix Copy. In 1978, Eric Schmidt developed a small network program for BSD while working at Berkeley on his master's thesis (Stevens, 1990, 9). His networking program was dubbed "Berknet" and was released along with 2BSD and provided support for up to 26 different hosts. Berknet provided support for up to 26 hosts and networking capability for BSD systems and included capabilities for transferring files and sending e-mail.

By the fall of 1980, Unix was able to run on two hardware platforms: the DEC PDP 11 and VAX 11 platforms.

In addition, it was able to be networked and operate as a networked system. All of this caught the eye of the U.S. Defense Advanced Research Projects Agency (DARPA), an agency of the Department of Defense (DoD), which was interested in seeing an its new Transmission Control and Internet Protocols (TCP/IP) running on multiple hardware solutions. The DoD had pretty much decided on the VAX as its primary platform and was in search of a suitable operating system.

Among its desires was an operating system that supported virtual memory (the ability to use the disk as a virtual memory image, thus effectively significantly expanding available memory). VMS (Virtual Memory System—the native operating system for the VAX produced by DEC itself) was an obvious choice. So was BSD Unix. In addition, because Unix was written in a portable language such as C, Unix had the advantage of being able to move to new hardware platforms, beyond the VAX, as they became available. This, along with the availability of source code, another Unix invention, sold DARPA on Unix. DARPA began to fund the research effort of BSD, and through this funding, BSD version 4.x (4BSD) became the first Unix to offer the TCP protocol running on the new Ethernet hardware, another interest of DARPA.

In September 1983, 4.2BSD was released officially, and it included TCP/IP networking and the Berkeley Socket interface. Sockets form the core of the Unix networking implementation of TCP/IP, and allow different computers to publish "ports" to which other processes on other computers can "bind," allowing them to talk to one another over the common Unix file abstraction via read() and write(). This release is most significant because it defined what Unix networking was to look like, as well as fundamentally delivering on DARPAs commission. BSD's socket architecture has become a networking standard and has been incorporated in most modern operating systems, including Microsoft Windows. The Berkeley socket interface now forms the fundamental network connectivity for all Internet traffic.

BABEL: The Commercialization of Unix

In 1982, the Justice Department concluded its decades-long dispute with AT&T by agreeing to a landmark agreement: the dissolution of Western Electric entirely and the divestiture of AT&T's phone service into the now famously independent "Baby Bells." With the Baby Bells handling the phone service, AT&T was now free to engage in business activities entirely outside of its traditional albatross: the pure delivery of phone service. This meant they could begin to sell software, including Unix, a plan they put into action immediately under the auspices of a new corporation called AT&T Information Systems.

This also meant that AT&T was considering its source code to be intellectual property and had to protect its distribution, lest it lose its status as a trade secret (which, ironically, it ultimately did). This meant that John Lions's wonderful commentaries on the Version 6 source code (he had transcribed thousands of lines of pure source in the commentaries) were now illegal, and their publication and distribution was banned. For decades, his commentaries were photocopied over and over again and distributed underground. Having in hand one of these "original" photocopies is a prized possession among current Unix cultural aficionados.

The net effect of the 1982 decree was that Unix became a commercial entity, which resulted in a proliferation of vendor-based Unix offerings. AT&T sold its own versions that derived from Version 7, with corporate license prices rising as high as $200,000 for a single source license by 1993 (Salus, 1994, p. 222).

By June 1982, work on 4.2BSD had already begun, and Bill Joy saw the commercial writing on the wall and left the CSRG at Berkeley to join with Andy Bechtolsheim, who had designed a new workstation capable of running 4.2BSD. Their new company was called Sun Microsystems, and their new workstation was dubbed the Sun Workstation. Sun took the 4.2BSD core and began to sell its own SunOS, based on 4.xBSD, on its workstations.

Other companies followed and entered the competitive foray. IBM came out with AIX (AT&T flavor) for its workstations, Hewlett-Packard offered HP-UX (AT&T flavor), Digital came out with ULTRIX and then later Digital Unix (BSD flavor), and Silicon Graphics offered IRIX (AT&T flavor). The list of commercial offerings exploded, with a multitude of incompatible Unix systems and new licensing costs.

The issues in licensing, pricing, and portability prompted an exasperated reaction in the academic community, used to a Unix tradition of open source access, and this reaction prompted several significant results. The first was a move by the Institute of Electrical and Electronics Engineers (IEEE) to form a committee in 1986 to define a formal standard for Unix-based operating systems to try and regain some compatibility. This standard was called POSIX, which stands for Portable Operating Systems based on Unix.

The second result was an announcement by AT&T, in 1987, to purchase some shares in Sun Microsystems, which resulted in a powerful marriage between the BSD and AT&T versions. The first major result of this new relationship was the integration of SVR3 and 4BSD into a new operating system offering from Sun called Solaris. The second result was the creation of the Open Software Foundation (OSF), a group formed by the "other" Unix vendors, who did not want to be left out in the cold. This group included Digital, IBM, and Hewlett-Packard as heavy hitters, among others.

The commercialization of Unix, and the concomitant cost of licenses and end of free source code, led to several significant free alternatives. Foremost among these is Richard Stallman's creation of the Free Software Foundation (FSF) in 1984. Stallman was working in MIT's Artificial Intelligence Laboratory and was offended by all the commercialization and licensing issues of the new proprietary Unix versions. The FSF is responsible for delivering to the entire Unix community the Emacs editor, which Stallman himself wrote, as well as a whole suite of tools that were "free"—free not in the sense that they did not cost money but that they were *free* to be shared, improved, enhanced, everything provided with source code but never to be commercialized. The software from Stallman's FSF was dubbed "GNU" software. GNU is a recursive acronym meaning "Gnu's Not Unix," for Unix had become in

Stallman's view a four-letter word in its commercial incarnation and licensing practice. Under Stallman's GNU "Copyleft" license, users can do just about anything they want with the source code, except prohibit its further distribution in source code form. It was called a Copyleft license because it did the opposite of the usual Copyright: It limited *not* the redistribution of the code, but rather, any restriction of its redistribution. The GNU Copyleft license makes the code and the license inseparable and therefore free.

The FSF has put out under the GNU license a prodigious amount of software and has actually made the open source movement what it is today. These GNU offerings include compilers, shells, games, debuggers, and a host of others. The GNU Copyleft license has also been incorporated by other efforts as well, making a wide range of free software available for a variety of Unix platforms, commercial and free.

The commercialization of Unix is also what prompted people like Andy Tanenbaum to develop Minix, a free version of Unix, written from scratch without a single line of AT&T code. It is also what sparked a young student in Helsinki, Finland, Linus Torvalds, to write his own Unix-like kernel dubbed Linux in 1990. The ubiquitous success of Linux has been nothing short of astounding. Linux, however, stands within the tradition of open-source philosophy, not as its instigator but rather just another participant in a long lineage. The open source of the original Unix and the subsequent movements that sprang up as a result of its commercialization are the true instigators of the open source movement.

HOW TO RECOGNIZE A UNIX SYSTEM WHEN YOU SEE ONE

What Is a Unix System?

A Unix system is an operating system that is made up of two environments, one, called the "kernel" and the other called the "user environment." The kernel provides low-level support for the user environment and interfaces the user environment with the physical hardware of the computer. Unix users operate within the user environment level and the kernel provides constant support for a user's commands by providing services that may be executed on behalf of a given user's request. Unix users interact with the computer through communication with a program known as a "command shell." It is the shell that most users associate with Unix.

Files and Directories

A Unix system provides users with the ability to store information on a disk, which provides for long-term storage capabilities. Data are stored in files, and files may be stored within directories, which are essentially composite abstractions for storing and arranging files into related or logical groups. That is to say, I might wish to store my personal files in a "personal" directory, my financial files in a directory called "money," and my work files in another directory called "work." In reality, however, directories are just files like almost everything else on a Unix system.

Files and directories provide Unix users with an easy way to name and arrange data that they (or the programs they run) need to save. Files are physically stored as a series (sometimes called a "stream") of bytes. Directories are named files that collect or organize other files. Files are stored on disks that are formatted as file systems. Both files and directories may be moved from one location (in a hierarchy) to another, renamed (in Unix, "moving" a file from one name to another). They may also be deleted and modified.

The Unix file is an abstraction that allows a user to put a name on information that he or she may wish to access again later. All the details of the physical location of the file on the file system on the physical disk, and other pieces of information necessary to the storage of that information, are hidden from the user (but always accessible to the curious) by the Unix file abstraction. Thus, users can create files by naming them and also open files, read and write data to them, and close them when they are finished.

System Users

Files do not exist on the system accidentally. They are owned and created by users of the system. Unix systems have users that operate in the user environment. Users have numbers associated with them known as "userids." Each user also has a "username," which is associated with their particular userid. The username is the name under which the user accesses the system. Users also have passwords that identify them as legitimate users of the system. A user can use her userid and password to gain access to the system by "logging on" to the system. All passwords are stored on the system in an encrypted form, never in plain text. No one can "see" your unencrypted password, even a system administrator.

Users may be arranged according to logical groupings. For instance, the users bob, jane, and linda may all be accountants, and therefore belong to the "accountants" group. The users bill, cindy, and ellen may be members of the "managers" group, and the users karen, david, ian, and evan might be all members of the "executives" group. Groups are used to organize users into various categories, allowing for the ability to assign certain access rights to groups of individuals rather that on an individual basis.

There is one special user on every Unix system, and that is the system administrator, sometimes called the "superuser" or simply "root." The superuser administers the system, which includes tasks such as managing the user community (creating accounts for users and managing groups), changing permissions on files, managing the system as a whole, scheduling system backups, installing software, and so forth. The superuser can do almost anything on a Unix system, so access to the superuser's password (sometime called the "root" password) is highly controlled. Anyone logged in as the superuser can, for example, read other users' mail, read their files (unless the user has performed additional steps such as encrypting their file data, etc.), modify other users' files, delete files, and so forth. A nefarious user who has obtained the root password could, if he or she desired, delete every file on the Unix system so that the system would no longer start up.

Login Processing

Because Unix is a multiuser environment, the system needs to provide a way for multiple users to use the system at the same time. This is done through a process known as the "login process." When a typical Unix system starts up or "boots" (known as "bootstrapping"), the masterboot sector of the primary hard disk partition is read into the computer's memory. It does several things, but eventually the unix kernel is started (a program on the disk called something like "unix" or "vmunix"). The kernel sets up the system, including initializing process tables, creating memory areas, and establishing certain buffers and caches. After the kernel has established a sane state, the kernel spawns (or "runs") a program known as "init." The init program is important because it ensures that the system is always ready in a particular "state." The Unix operating system has various states, including boot (initialization), restart, shutdown, single, and multiuser. The normal state is multiuser, and when init begins this state, another program known as a "getty" (spelled from "get" and "tty" which is itself short for "teletype," indicating a terminal) is started on a particular terminal. The getty process listens for activity from a terminal, and when a terminal is connected, getty runs another program known as "login," which prompts a user to log in to the Unix system.

It is the login program that the average user first sees as the interface to a Unix system. The user is prompted to enter a username. Once it has been entered, the user is prompted to enter his or her password. Once the user has entered the username and password, the login process examines these entries and compares the username and password (encrypted) entered by the user with the information that the superuser has entered for that user (users are usually allowed to alter their password within certain security constraints).

If the user has not entered the correct information, he or she is advised of the failure and is not allowed access to the system. At this point, the getty program spawns another log-in program and the user is prompted to log in again. If the user has entered the correct information, he or she is "logged in," and the log-in process spawns yet another program, known as a command "shell," which prompts the user to enter commands to the system. The shell interfaces the user with the rest of the Unix system and prompts the user to enter new commands (a Unix prompt usually is made up of a $ character or a % character, depending on which type of shell is being used).

The Home Directory

When a user is first logged on to a Unix system, the user's default directory is changed to a special directory known as the user's "home directory." A typical Unix system has as part of its directory hierarchy a directory called "home." Under this directory, all valid users on the system are given their own directory, so one might see something such as /home/bob and /home/sue, indicating both Bob and Sue's home directories. Users are usually given complete permission over their home directories, so that they can freely create new files there as needed. Users usually do their work in their home directories (or subdirectories off their home directory).

Online Help System

From the first release of AT&T Unix in November 1971, the Unix system has provided an online help system, called "manual pages." Manual pages, or "man" pages for short, are accessed by the "man" command. It is a cultural imperative that Unix commands come with their own man pages. For example, if I wanted to find out how to run the Unix "diff" command (which compares two files for similarities and differences), I would execute the following command:

```
man diff
```

I can also do a "keyword" search, by passing the -k flag to man, and so the command

```
man -k write
```

would list all of the commands in the Unix system that have something to do with "writing." Each man page details the name of the command, the syntax for running the command, and the various options for the command (options alter slightly the behavior of the command), and a general description of the command along with any problems that exist with the command (known as "bugs"). Part of the Unix culture was an honesty about its limitations, and from the beginning program writers have detailed in the man pages things that do not work quite right in their programs.

Pipes and Filters

One of the seminal influences on early Unix development, and one of the core reasons for Unix's eventual success, was the introduction of pipes and filters in AT&T System V3. The third edition saw the inclusion of Doug McIlroy's long-fought argument for the usefulness of macros. Prior to Unix, jobs were executed sequentially in batch mode, and Job 2 had to wait until Job 1 had been completed until it could begin processing. Essentially, Job 1 would write out its data into a file, which would then become input for Job 2, but only when Job 1 had completely finished writing the file. This prohibited concurrency, meaning that it was impossible for two jobs to be running simultaneously, both operating at the same time on a continuum of stream data. For a multitasking operating system, therefore, there needed to be a way for Job 1 to "send" its data to Job 2 *as it was processing it,* so that a stream of data might be considered to be *in process,* and Job 1 and Job 2 could be simultaneously working on different sections of the same data stream at the same time, with whatever Job 2 was working on having already been processed by Job 1. Strictly defined, this process was essentially handled by connecting the output of one program (Job 1) to the input of another program (Job 2). Thus, as Job 1 finishes processing some text, it hands that text off to Job 2, which immediately begins processing the data, even as Job 1 continues to work on the remaining input stream. This pattern came to be known as pipes and filters, with different jobs being known as filters, and the connection of input and output as pipes.

The net effect of this strategy was that on multiprocessor computers, jobs could be completed faster and programmed with much greater ease. Users working at the shell could benefit from pipes as well. For example, the command

```
ls -la/ usr/bin |  grep mark |  awk '{sum +
  = $5; print} END {print sum}'
```

will generate a detailed directory listing of the/usr/bin directory, and print out only mark's files (literally any line that contains the letters *m, a, r,* and *k* consecutively in that order) and then print out as the final output of the pipeline the total number of bytes in the selected files. It will do this by piping the output of the ls command into the grep command, and then piping the result of that stream out to the awk command. The filesize of a given file is stored in the 5th field ($5) of the long (-l) ls listing output. In the above example, grep and awk are both filters. The '|' character is the pipe command that creates a pipe between filters. Imagine asking the Microsoft Windows Explorer to provide you with that information on some arbitrary (i.e., user-defined) subset of files in a directory.

Using pipes allows the developer to write many small generic programs (called "filters") that focus on doing one thing well and then to be able to arrange them in various configurations that produce different and powerful user-defined results. Any program that is written to accept ASCII input and write ASCII output over a pipe is known as a filter.

CORE COMPONENTS

Editors

Unix has a long list of editors, beginning with Ken Thompson's original ed line editor. The two most popular editors on a Unix system are BSD's vi editor and Stallman's emacs editor. Other editors have come along in the meantime and are available on different Unix versions. Among these are the PICO editor (originated with the Pine mail program), nedit, and the various CDE-based editor versions, usually called dtpad or something similar. (CDE stands for the Common Desktop Environment and is a graphical version of Unix based on MIT's X Windows graphical user interface. CDE is available for most commercial Unix versions.) By far, the two most popular editors on a Unix system are vi and emacs. Every Unix system can be expected to offer the vi editor out of the box.

vi

Vi (pronounced "vee-eye") is a command mode editor, meaning that it is a shockingly far cry from the common WYSIWYG (What You See Is What You Get) word processors of today (and often much more powerful). Vi has two primary modes, command mode and edit mode. Users type their documents in an edit mode and modify them by issuing commands in command mode.

There are numerous commands to operate on text in vi, and they all operate from command mode. For example, to swap two characters (say "no"), the user would position the cursor on the n and then issue the command "xp" (x for cut and p for paste); the text would then read "on", with the two adjacent characters swapped. The mode commands include "i," which enters into insert mode; "o," which opens up a line under the cursor; "O," which opens a new line above the cursor; and "ESC," which exits an edit mode. Although for many users it is difficult to get used to vi, a competent vi user can completely transform a large file of text before a Windows notepad user has even navigated the Start menu, launched the notepad program, and highlighted the first bit of text with his mouse.

emacs

The emacs editor is radically different from vi in that it is not a modal editor but rather an editing environment. It is powerful and can be used as a programmer's editor supported by special modes for dozens of programming languages. These modes provide various capabilities, such as colorized syntax highlighting and pretty formatting that can be customized by the emacs user.

Emacs was originally written for Unix by James Gosling (of the Java Programming Language fame) and offers an environment in the sense that users can do many different things within the emacs editor. For example, users can of course edit text, but they can also read and write e-mail, read and post messages through usenet, navigate and display directories and files in the file system—they can even play games with intriguing names such as "Zippy the Pinhead" and "Doctor" (a "psychiatrist" who will diagnose all one's mental problems—free of charge). You can do so many things from *within* emacs that before windowing environments many Unix users lived their entire day within emacs, rarely exiting out to the shell.

Command Interface: The Shell

Unix interfaces with the user through a program called the shell. A shell is a command processor that prompts the user for input and interprets and then executes the command entered on the user's behalf. A shell is just another user program. It provides the main interface between the Unix system and the user. As with most things in Unix, the user has a number of shells from which to choose. A user can choose to work within the Bourne shell (sh), the Korn shell (ksh), the C shell (csh), and a variety of newer shells, including the z shell (zsh), the Bash shell (Bourne Again Shell, bash), and the modern C shell (tcsh).

Different shells have different capabilities, but most shells draw from two primary heritages: the original Bourne shell from AT&T Version 7 and the original C shell from BSD (see Figure 1). System V derivatives have usually followed the Bourne shell lineage, and most modern Unix systems come with the Korn shell or original Bourne shell as the default user shell. Users can generally change their shell preference, however, and the chsh (change shell) command will allow users to select another one. The traditional prompt for Bourne derivatives is the dollar sign: $. The traditional prompt for C Shell derivatives is the percent sign: '%'. This will also be an indication of what type of shell a user is running.

Because a shell is just another user program, from one shell, a user can start up another shell. A user can run a C shell session from a Korn shell, for example. A user can

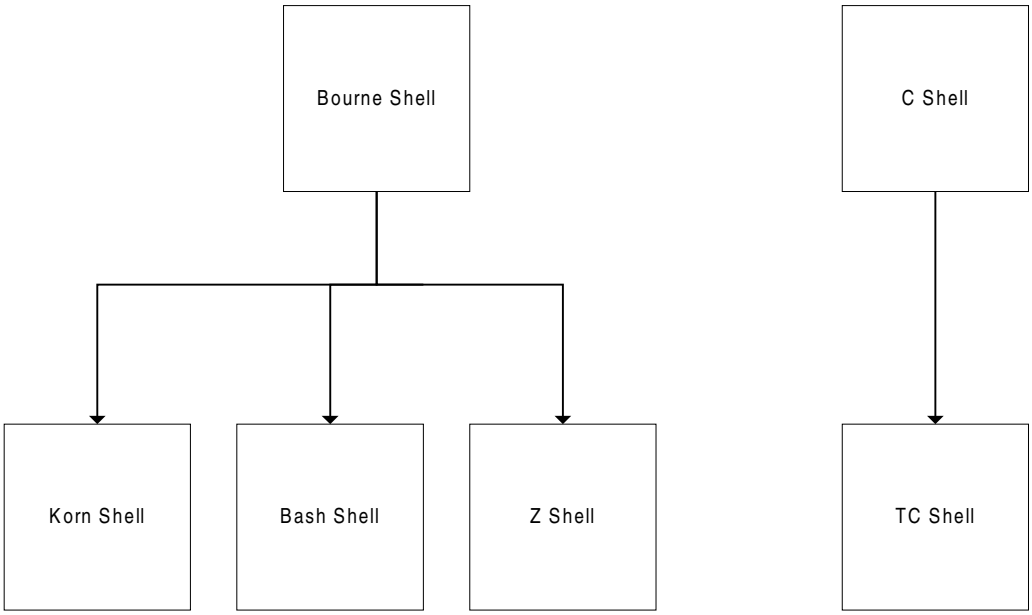


Figure 1: Shell lineage.

determine what shell her or she is currently running by issuing a command such as

```
echo $SHELL
```

The above command will print out the user's default shell, something like "/bin/sh," indicating the original Bourne shell, or "/usr/local/bin/bash," indicating the Bash shell. An example of this command is given in Figure 2.

Although different shells support different features, several significant features generally appear in one form or another. Among these features are redirection, command history, command line editing, command line completion, and aliases.

Entering Commands

Users can enter commands to run programs at the shell's command prompt. Every unix command has a command name, which the user can use to run the program. For example, the program to copy one file to another is called "cp." To execute the copy command, the user would enter

```
cp file1 file2
```

which would copy the contents of file1 into another file called file2 (creating that file if necessary).

The execution of commands can be tailored by the user by issuing defined options (called "flags") or supplying arguments to the command. Options are preceded by a single dash character '-'. For example, the standard command to list out the files in a directory is the "ls" command. By default, the "ls" command with no options will just list the names of the files in the directory. By giving the '-l' flag or option, however, more detail is give about each file, including the files owner, group, access permissions, file size, number of links, and so on.

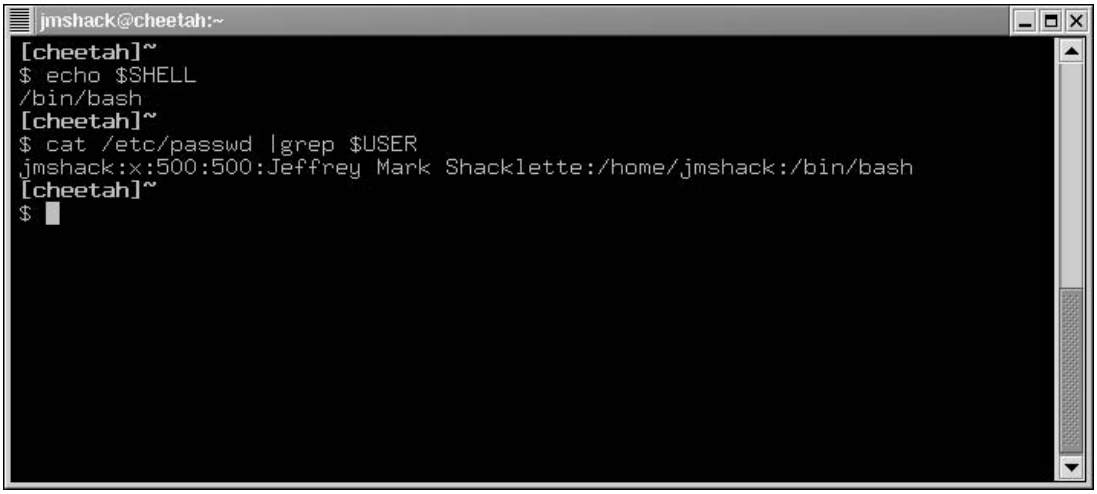


Figure 2: Discovering a default login shell.

Giving the '-a' flag will list *all* files, including certain hidden files that are not normally shown by default. Giving the '-1' flag will list filenames all in a single column at the left of the terminal screen. Giving the '-i' flag will list the file's inode number next to its filename. Giving the '-R' flag will recurse through subdirectories, list out the files in those directories as well as the current directory (default). Giving the '-S' flag will sort the files according to file size. Giving the '-Sr' flag will sort the files in reverse order.

Options vary according to the flavor of Unix. For instance, certain options apply to AT&T derivatives, and other options apply to Berkeley derivatives. That is to say, a given command (most notoriously the 'ps' command to list out all running processes) have different flags based on whether the ps command is on a Berkeley-based or AT&T-based Unix. Nevertheless, options can allow the user to tailor the standard output of a Unix command to suit their special interests or needs.

Redirection

By default, output from a command (if there is any) goes to a special file called standard output. By default, standard output is set to the current user's terminal (also a file). So if I issue the command "date," the current date will be displayed on my terminal. Because everything is a file in Unix, it is easy to redirect the output from one file to another, that is, from the terminal device to some other file. For example, the command "echo hello >hello.txt" would redirect the standard output of the "echo hello" command into a new file called "hello.txt," effectively creating the file called "hello.txt" (if the file already existed, its contents would be overwritten with the redirected output). Thus, if I were to type out the contents of the new file hello.txt, I would see that its contents was "hello":

```
$ echo hello >hello.txt
$ cat hello.txt
hello
```

Unix assigns three default files for use on the system: the "standard input," "standard output," and "standard error" files. All files are designated in the system by something known as a "file handle," and our default files have three associated file handles. Standard input has file handle number 0, standard output has default file handle number 1, and standard error has a default file handle number 2. These file handles allow users to reference them during redirection.

Standard input is a file through which the system delivers data entered to the program by either the user interacting with the shell or by some other program through a pipe. When a program wishes to write text out to the user's terminal, the program's author can choose which of the two "output" files to write to, either standard output or standard error. Generally, when a program's author wants to display information that is part of the general operation of the program (often the information the user wants to see by running the program in the first place), that information is written "to" the standard output file, which is by default sent to the user's terminal screen. If the program's author wants to write out some type of error-related information, the author will write the error information out to standard error. By using this standard, the user can enter (in Bourne-derived shells) the following syntax:

```
mycommand 1>standard.output 2>errors.txt
```

and expect that the program's main information would be redirected out to a file called "standard.output" and that errors, if any, would be written out to a separate file called "errors.txt." This allows a user to "declutter" the output they are interested in from the various errors that, though not necessarily critical, still might be distracting.

Redirection is one of the many concepts in Unix that was adopted by subsequent operating systems (although not to the same degree of depth), in limited form, such as MSDOS and later Windows NT.

Command History

Most Unix shells maintain a list of previous commands entered at the command prompt. These shells also provide some means for accessing previous commands and parts of previous commands. Although the precise shell syntax varies, command history serves one main purpose: to reduce typing at the command line. Say a user wants to use elm to mail the same e-mail message to a number of users, he or she types the following commands at the prompt, successively:

```
elm -s "please come to my Halloween party
  tonight" linda <mail.invitation.txt
elm -s "please come to my Halloween party
  tonight" ellen <mail.invitation.txt
elm -s "please come to my Halloween party
  tonight" steve <mail.invitation.txt
```

Now, even in these few commands, there is a significant degree of repetition. The user is sending the same e-mail message mail.invitation.txt (via redirecting input) with the same subject to three people. Notice also that every command ends with the filename mail.invitation.txt.

Now, the user could of course simply type this redundant text over and over again. But what if she were throwing a large party and needed to send out 50 invitations? That's a lot of typing. The command history mechanism offers a number of shortcuts at the command line that save us from having to retype repetitive data. For example, the lines could also be executed (with the tcsh as well as the bash shells) as follows:

```
elm -s "please come to my Halloween party
  tonight" linda <mail.invitation.txt
^linda^ellen
^ellen^steve
```

Clearly, a great deal less typing at the expense of some really strange looking syntax. Although I cannot go into the details of command line history, suffice it to say that the command syntax *^search^ replace* will search through the previous command and replace *search* with *replace*. So, once the command "vi myfile1.txt" has been executed, immediately executing the command "^1^2" would execute the command "vi myfile2.txt," effectively replacing the 1 from the previous command with a 2.

Table 1 Command Line Shortcuts Available in Shells Supporting Command History

!:0	The previous command name
!:*n*	Repeat argument *n* of the previous command
!:^[or !^, leaving out the colon]	Repeat the first argument of the previous command
!:$ [or !$, leaving out the colon]	Repeat the last argument of the previous command
!:*n1–n2*	Repeat the arguments between *n1* and *n2* from the previous command, inclusive
!:-*n*	Repeat arguments 0 through *n* from the previous command

Table 1 lists some of the command line shortcuts that are available in shells supporting command history.

Command Line Editing

Most modern Unix shells (with the exception of the original Bourne shell) support a feature known as command line editing. Command line editing allows one to use the history buffer to move back to previous commands and modify them using a line editing syntax from Unix editors, namely, vi and emacs. The user can also edit current command lines as he or she types. For example, in the Korn shell, a user can tell the shell that he wants to edit command lines as if he were in the emacs editor, by issuing the command "set -o emacs" ("set -o vi" would set the editing for vi emulation). Once he has set emacs emulation, the user can navigate the command history and command line using familiar emacs cursor movement and editing syntax.

For example, to go to the previous line in an emacs buffer, one would press Ctrl-p. To go to the next (subsequent) line, one would press Ctrl-n. To move backward on a line, character by character, one would press Ctrl-b repeatedly. To move forward character by character, one would press Ctrl-f. A user can walk backward character by character on a command line, by pressing Ctrl-b. The ability to search backward through the command history is also available in many shells through the Ctrl-r key combination. Table 2 shows some of the more common emacs command line editing commands.

Command Line Completion

Coupled with command line editing is command line completion. Imagine that there are three files in the current directory: file1000a, file2000b, and file2010c. A shell that offers command line completion will allow the user to type just enough of a filename and then press a command line completion key, and the shell will examine the file system and *complete* the rest of the filename automatically. For example, suppose a user is using the bash shell, enters the command "ls file1," and then presses the tab key. The tab key is the completion key for bash and tcsh (it's ESC ESC in the Korn shell). The shell automatically completes the filename, leaving you with "ls file1000a" on the command line.

Table 2 Common Emacs Command Line Editing Commands

Meta-b	Move one word back on the command line (the Meta key is often the "Alt" key on a PC keyboard)
Meta-f	Move one word forward on the command line
Meta-backspace	Delete the word directly before the cursor's current position on the command line
Ctrl-a	Move to the beginning of the command line
Ctrl-e	Move to the end of the command line
Meta-d	Delete the next word in the command line

Aliases

Sometimes users want shorter names for lengthy common command lines. Suppose, for instance, that every time a user wanted to run the emacs editor to edit a file, she wanted to set the colors and font size. So for every file she edits, she would need to type something like the following line:

```
/usr/bin/emacs -fn 6x13 -fg black -bg cyan
  -cr blue -ms white filename.txt
```

The user would certainly get tired of typing that long command every time she wanted to edit a file. Aliases can be used to store commands for later use. For example, one could store the core of the previous command (minus the filename) in an alias, and then simply use the alias to refer to the longer command:

```
alias myemacs = "/usr/bin/emacs -fn 6x13 -
  fg black -bg cyan -cr blue -ms white"
```

Once that command is issued, the user could then simply type "myemacs filename.txt," and the myemacs alias would expand into the longer command line associated with the alias, setting all her desired colors and the font size to 6 × 13. Aliases, and other shell commands (such as the set -o emacs command) are generally placed in startup command files, which are default shell scripts loaded by the shell each time a user logs in. In bash, the main file is .bashrc. In the Korn shell, it is.kshrc. For the C Shell, it is .cshrc (.tcshrc for the TC shell).

Shell Variables

Shells also allow the creation of user-defined variables that enable the shell programmer to store temporary values. A shell variable is usually an all capital letter name referencing a particular value. Shell variables are accessed by placing a dollar sign in front of the variable name. For example, one of the standard shell variables is called $USER, which contains the username of the user currently logged into a given terminal. The echo command will print out to the terminal the value of any variable. So, issuing the

command "echo $USER" might print out "linda," if the user with the username linda happens to be currently logged in. Other common default shell variables include $HOME, which indicates the current user's home directory within the file system. The command "cd $HOME" will change the current directory to the user's home directory, for instance. The variable $PWD stores the present working directory, or the current directory the user happens to be in at the moment. The $PS1 variable contains the prompt that the user can set. For example, the following Bash command will set the prompt to display the host machine name followed by the current date and time before the default dollar sign:

```
export $PS1="[$HOST][`date']$"
```

This would produce a prompt that might look like this:

```
[computer1.cs.uchicago.edu][ Sun Mar 30
  17:45:54 2003]$
```

The introduction of the word 'date' in backward quotes (often called "backticks") introduces a nice capability of some Unix shells: the notion of *command substitution*.

Command Substitution

Command substitution exists to allow one command to be included within, and executed before, another command. This is a powerful feature. Let's say that a user wanted to edit all the files containing the word "Unix" in a particular directory. He could obtain a list of the files with the word Unix in them by issuing a grep command in the form of "grep -l Unix *." This command uses the "*" wildcard to indicate "all files." Grep is a command that will search through a file or files and look for a particular regular expression. The regular expression in our example simply states that we are looking for the letters *U, n, i, x,* in all the files in the current directory. The -l flag will indicate to the grep command that it should only print out the *name* of the files that match. Let's say that we had four files that contained the word "Unix": file1, file2, file3, and file4. If we wanted to edit all of these files with the word "Unix" in them, we could issue the command

```
vi `grep -l Unix *'
```

which the shell would automatically expand into the following command before being executed:

```
vi file1 file2 file3 file4
```

This would in effect launch vi with these four files, and the user could do whatever edits he wanted to those selected files, effectively editing all files in the current directory that contained the word "Unix."

Shell Scripts

Every Unix shell supports a shell programming language. Shell programs are commonly called shell scripts. Shell scripts are simply lists of commands that are executed by the shell interpreter one right after another. Any command that can be called from the shell prompt can be added to a shell script and vice versa.

A shell script can call another shell script, so a small program can be created of multiple coordinating scripts. Some of the Bourne shell derivatives (e.g., Korn, Bash) also support the concept of shell functions, which are loaded once (usually at login) and generally execute more quickly than shell scripts. Functions also allow the shell programmer to give a more structured approach to shell programming, where one shell script may call other procedures (scripts or functions).

Shell scripts can be fairly powerful tools, and they are highly portable. Unlike binary programs, which must be recompiled to run on different Unix systems, shell scripts, because they're written in that common Lingua Franca known as text, can generally be run on any other Unix system that provides the same shell.

The shell scripting support provides some default language capabilities. I have already mentioned the ability to create shell variables to store data. Shell scripting languages provide flow control syntax (if. . . then. . . else), including for loops, while loops, and conditional clauses in the form of if-then-else syntax. The shell scripting language also includes certain string comparison operations, numeric comparison operations, and some default type conversion capabilities, along with the ability to walk through a command line and operate on the various options (preceded by a dash in Unix) and command line arguments (arguments passed in to the shell script at runtime).

Job Control

In addition to incorporating a shell scripting language, shell variables, and command substitution, most Unix shells (with the exception of the Bourne shell) support something that Bill Joy added to his original C Shell: job control. Unix is a multitasking operating system, allowing a user to run multiple commands at the same time, but originally, if a user wanted to execute three commands simultaneously, he would have to open three terminal sessions, log in three times, and then execute a command in each of the three terminals. Bill Joy thought it would be nice if within a single terminal session, multiple commands could be executed, and be able to run in the background, while the user continued to work at the prompt.

Job control allows a user to (a) execute a job (process) in the background, (b) move a background job to the foreground and vice versa, (c) temporarily suspend a background or foreground job, (d) stop (permanently kill) a background job, and finally (e) get a list of all jobs that are currently running in the current shell. To start a program in the background, the user simply follows the command with an ampersand character (&). When a job is placed in the background by the shell, the shell automatically assigns that job a "job number." The user can subsequently use the job number to reference the various jobs running in the background. The user can get a list of all the various jobs running within a single shell by executing the "jobs" command. Figure 3 shows the jobs command output with four jobs running: an emacs session, xclock, the user's fincalc program, and the user's vi session for a term paper. A list of common job control commands is given in Table 3.

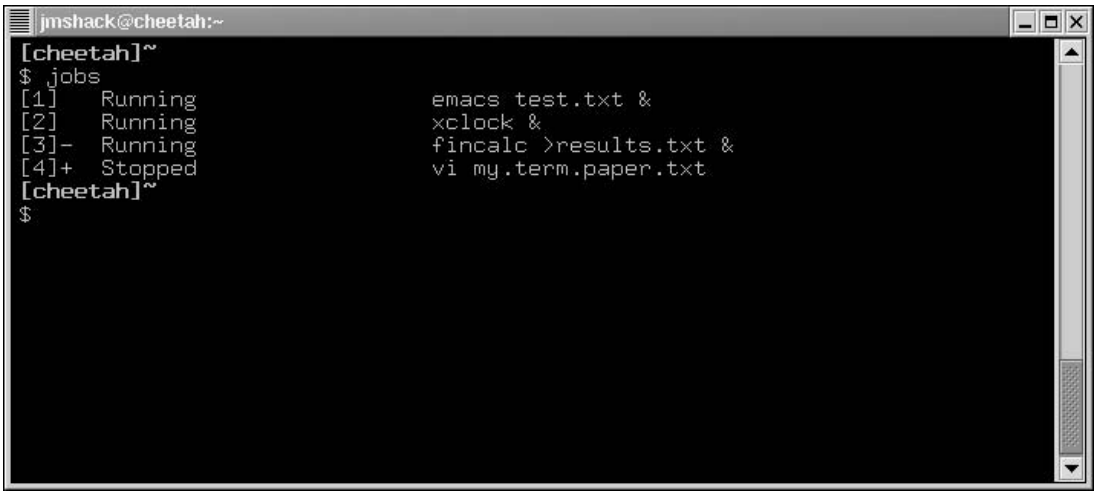


Figure 3: Job control.

Summary of Popular Commands

There are a number of commonly used commands on a typical Unix operating system. For a detailed list of available commands, users should visit any number of standard texts on Unix (most notably Gilly, 1986; Hahn, 1994; Peek, Todino, & Strang, 1998; Sobell, 1995). Online resources include The Unix Reference Desk (n.d.) and Digital Unix Online Documentation., which details Digital Unix (a BSD-flavor) but is highly useful for the standard commands and syntax. Table 4 summarizes some of the most popular commands.

Table 4 Popular Unix Commands

cat	A filter program that prints out a file to standard output.
cd	A utility to change the current directory to some other directory in the file system
chmod	A command to change the permissions of a file
cp	A utility that copies one file to another
lp, lpr	Commands to print files to a printer
ls	A utility for listing the contents of a directory on the file system
man	Access the help system for information on commands and functions
mkdir	A utility for creating new directories
mv	A utility to move a file from one location to another or to rename a file
rm	A utility that removes a file from the file system.

OPERATING SYSTEM STRUCTURE

Kernel Structure

The job of an operating system is to provide an interface between the user of the system and the hardware of the computer. This interface is provided, at the lowest level, by the part of the Unix operating system known as the *kernel*. The kernel is a file that exists on the file system like any other file but is read into memory when the computer is first turned on (a process known as "booting"). When the computer is booted, the kernel file (usually called something like "vmunix" or simply "unix") is loaded into memory and begins executing.

The kernel provides low-level services and runs on the computer in a protected mode, which tries to ensure that that the operating system itself will always be available and in a "sane" mode for users to use. Included among these low-level services provided by the Unix kernel are terminal handling services, signal handling, user management, process management and scheduling, memory management, file system services, hardware interrupt services, and interprocess communication.

These low-level services are provided to both the kernel itself and to user processes by a defined set of function calls known as an API (application programming interface). The API provided by the kernel offers functions called "system calls" that can be called by a user's program, and these calls provide fundamental access to services such as memory management, file system access, and so on. Figure 4 shows an abstract view of a typical Unix kernel layered architecture.

Table 3 Job Control Commands

Ctrl-z	Suspend the job in the foreground
Ctrl-c	Terminate (kill) the job in the foreground
fg%*n*	Change the background job number *n* to run in the foreground
bg%*n*	Change the stopped job number *n* so that it begins executing in the background
kill%*n*	Terminate the job number *n*
jobs	Print out a list of all current jobs known to the current shell

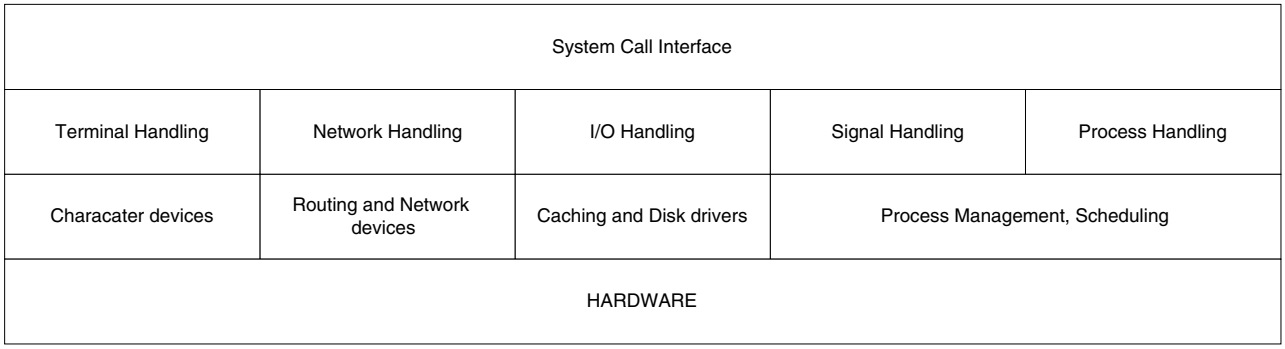


Figure 4: Structure of a typical Unix kernel.

Filesystem

The typical Unix file system is a hierarchical composite of directories and files, starting with a base directory called root. Every directory in the system, with the exception of the root directory, belongs to some other directory. Every regular file in the system is contained in a single directory. Figure 5 shows a typical layout of a hierarchical Unix file system.

The main directories and their functions on a typical Unix system are given in Table 5.

The /home directory is significant because most users on a Unix system are assigned a default home directory, and this directory is generally stored under the /home directory. Therefore, a user with the username andy would usually have a home directory under /home/andy. The user sally would have a home directory under /home/sally. A user's home directory is the current working directory immediately after log-in and is the place where the user can store personal work files.

Each file in the system is associated with a structure known as an inode. An inode (information node) contains information relative to that particular file in the file system. The inode contains, among other things, the filename, the current size on disk in bytes, the type of file (regular file, directory, link, device, etc.), the number of hard links to the file, information indicating the user and group associated with the file, a file system unique identifier, and some time stamps representing when the file was last read, modified, and when the inode structure itself was last changed. Most important, the inode contains a pointer to the physical location of the file's data on the actual hard disk.

Every file in the file system can be defined as being somehow derivatively off of the root directory. A file with a pathname that begins with a / slash is called an absolute pathname, because its location within the logical hierarchical structure is unambiguous. An absolute pathname describes exactly how to navigate to a particular file on the system. For example, the filename/etc/security/dev/audio is an absolute pathname because it denotes a single file on the system, unambiguously, starting from the root directory.

A relative pathname is a pathname that begins from the current working directory. Relative pathnames often begin with two periods (representing the parent directory of the current directory) or adirectory name in the current

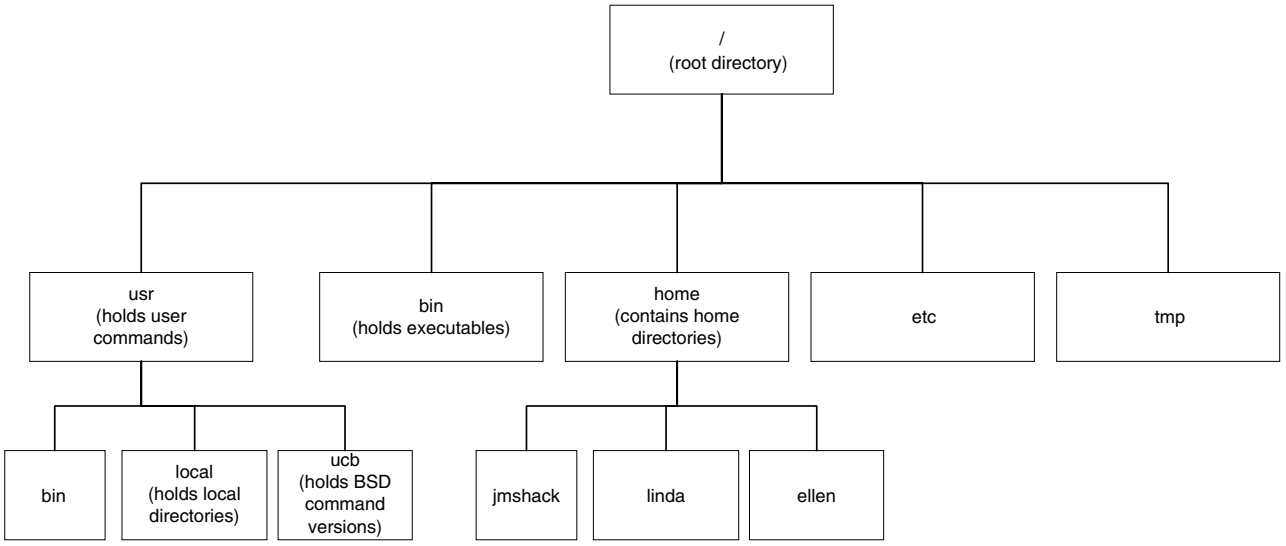


Figure 5: A typical Unix files system layout.

Table 5 Main Directories and Their Functions on a Typical Unix System

/(root)	The base directory of all other directories
/bin	The directory that holds system binaries
/usr	The directory that contains user commands
/etc	The directory that contains system configuration files, including the files involved in system initialization
/tmp	A world-writable directory for storing temporary files (deleted periodically)
/home	The base directory for all user's home directories

working directory. Such pathnames are called relative because they are completely relative to whatever directory the user happens to be in when the relative pathname is referenced. For example, if the current working directory happens to be the root directory itself, the absolute pathname /etc/security/dev/audio denotes the same file as the relative pathname etc/security/dev/audio (notice the missing initial '/' slash). If the current working directory is a sibling directory (a directory at the same level) as /etc., for example, /bin, then the audio file could be represented by the relative pathname "../etc/security/dev/audio." Again, the pathname is relative because (a) its validity depends on the current working directory and (b) the pathname does not begin with a '/' slash, denoting the root directory.

For security, Unix files may be assigned permissions as to which users and which groups of users may access them, and in what ways. For example, I might create a file in a directory called "personal.stuff" and name that file "confidential.info." I might decide that only I can modify that data, but I might wish to allow other executives to read (but not modify) the file's contents. I might also decide that no one else should have access to that file in any way, effectively disallowing others from using the file in any way. In Unix, access to files and directories can be controlled through the use of file and directory permissions.

Security

Security on a Unix system is a layered approach defined in terms of access to the system itself and its resources, including files and directories. Access to the system is managed by encrypted passwords associated with userids during the login process.

Files may be read from, written to, and executed, depending on individual permissions. Every file in the Unix file system (remember directories are files) has a particular user who "owns" the file, as well as a group that is associated with the file. Groups collect users and give the ability to assign rights to a file to a group of users in addition to the single user who owns the file. One can see the file permissions for most any file in the system by typing the "ls -l" command. For example, if there is have a file called "inventory.stat" owned by the user "tom" and associated with the group "inventory," one might see a long directory listing such as the following:

```
-rw-rw-r-- 1 tom inventory 1098764 May 1
  2001 inventory.stat
```

This set of information tells us the following:

The file's name is "inventory.stat";
The user "tom" owns the file;
The file is associated with the group "inventory";
The file contains 1,098,764 bytes;
The file was last modified on May 1st of 2001;
There is 1 hard link to the file; and
The file's access permissions.

According to the symbolic rights of the inventory.stat file, the permissions are represented as "rw-rw-r- -." This means that the owner would be able to read and write the file (the first "rw-" in the triad), the members of the group "inventory" associated with the file would also be able to read and write the file (the second "rw-" in the triad), but all others would only be able to read the file (the final "r- -" in the triad).

THE UNIX PHILOSOPHY

Write Small Programs That Do One Thing Well

From the beginning, Unix has had a penchant for small programs. Some of this is due to the inherent memory limitations of the early Unix systems, but another more important reason was the philosophy that programs should be small, do one thing, and be written to act as filters. Thus, many small programs could be linked together using pipes to create a much larger capability. Unix users can create their own "super commands" by putting together numerous filters in a new order to accomplish some larger activity.

Other operating systems provide large commands that do a lot of work but that are not easy to tailor. By providing a number of small programs that do a number of specific things, but that can be arranged in succession to accomplish much larger scale tasks, Unix offers its users both the advantage of power as well as flexibility—an addictive combination.

Write Programs to Communicate Over a Common Data Format

Integral to the concept of pipes and filters is the emphasis on writing data out in a common format. The key to being able to "plug and play" various filters is that filters are designed not to accept some proprietary data format, but are designed to work generically with standard ASCII text. By forcing a constraint that all filters must deal with plain ASCII text, filters can use the generic pipes that Unix provides to communicate with one another without incurring an additional overhead by having to marshal and demarshal incoming and outgoing data. So Unix can be said to have a penchant for plain text, and this is partially why all Unix editors (ed, emacs, vi, ex, pico, nedit, etc.) work with plain ASCII text. Contrast this with a data file created by a proprietary program such as Microsoft Word. In that format, the only program that can successfully read that file is Microsoft Word. No other tool can be leveraged to work on that file, implying that if the

functionality you need is not in Microsoft Word, you are out of luck.

Now certainly the decision to use a common textual stream format has performance implications, but another fundamental philosophical tenet of the original designers is that Unix is predisposed to favor interoperability and portability over raw speed. It was assumed that faster hardware could always be deployed to speed up two programs that were communicating but that nothing could be done without serious labor to help two extremely fast programs that couldn't communicate because they each dealt with mutually incompatible data formats. So interoperability is preferred as a fundamental principle over speed.

Everything Is a File

One could almost say that one effect of the hierarchical file system concept was that the designers of Unix liked files so much that they decided to make everything a file. When Brian Kernighan and Rob Pike published their landmark Unix programming book, *The Unix Programming Environment,* the first sentence of the chapter on the file system simply stated the following: "Everything in the UNIX system is a file" (Kernighan & Pike, 1984, 41). It is not so much that every thing in Unix is a file, but that so many devices in Unix conform to the same abstract pattern; they follow the same, consistent API. That is, whether a user is opening a regular file on a disk, writing to a tape drive, reading characters from a modem, printing something to a printer, writing out to a pipe, writing out to a socket, or doing just about anything else, the process implies reading and writing bytes of data. This means they are using the same core set of system functions to access the file: open(), creat(), read(), write(), and close(). As far as the programmer is concerned, each of these types of "files" is just another stream of bytes to be read from and written to.

The device abstraction isolates the programmer from the details of how a particular devices actually works. So through the standard file system calls, the programmer is relieved from the details of knowing how to advance a tape physically, how to communicate with a serial port, and so on. These details are hidden from the programmer within the respective drivers in the operating system, but these devices are all accessible behind a common and consistent interface. This sort of consistency also aided Unix's adoption, promotion, and success.

Central Priorities

Portability

Writing to a common data format brings up the emphasis on portability. One of the key problems with operating systems before Multics was that they tended to be written in a machine-dependent assembly language. Porting operating systems from one hardware platform to another was very labor intensive. With the Multics and Unix emphasis on writing programs (including the operating system itself) in a portable language, the benefits of reuse became apparent. Source code could now be recompiled on another computer with an assembler program, and with (usually) a relative minimum of fuss, that program would now run on the new system. One of the primary reasons DARPA adopted BSD Unix over DEC's proprietary VMS operating system was that Unix was portable to future hardware platforms, whereas VMS was written specifically for the VAX and would have to therefore be rewritten in assembler for every new hardware platform to come along.

Simplicity

There is a certain parsimoniousness about Unix. It is evidenced in a new user's complaints that no output is generated from a command when it succeeds. The reasons for this are several. First, if the line "Congratulations, your command succeeded" is printed out by a program, it cannot therefore function as a filter because the "success" message would be output rather than the modified input stream. It will always print out the congratulatory message. Second, Unix was initially developed on teletype terminals (the device representing a terminal in Unix is tty, a unixism for TeleTYpe), that could print out some 30-odd characters a second, certainly slow by today's standards, but not too cumbersome in the 1960s. At least user could see the interaction between themselves and the computer. Nevertheless, teletype terminals were slow, and any way to minimize the unnecessary writing of output was to be preferred. Thus, generic success messages were to be avoided. This also explains, in part, why so many Unix commands are only one or two characters in length: wc, ls, db, du, ed, ld, ln, mv, ar, as, b, to name a few. The fewer characters to type, the faster the command could be executed.

Driven from its Multics and PDP 7 roots, Unix has always had a penchant for simplicity, which can be seen in the earliest releases of Unix. The sixth edition of the Unix operating system had less than 10,000 lines of code in it. By contrast, Microsoft's Windows NT 5.0 operating system is reputed to have more than 40 million lines of source code. Even now, with modern commercial Unix systems available, the source code lines still do not approach anywhere near that number. Never has there been an attempt to write a single Unix program that attempted to be all things to all people.

Elegance

One of the definitions for elegance in Webster's *New World Dictionary* is that of "restrained opulence." Oxymorons can often be enlightening, and this one is no exception. There is something restrained and nevertheless opulent about Unix. The restraint comes in its fundamental principles of smallness and simplicity. Yet there is a certain wealth afforded to the Unix practitioner, and this comes in the wealth of opportunity those small utilities offer. This wealth drives from the openness of the operating system itself. Because of the toolbox practice, larger tools can always be constructed from smaller ones, and the essence of Unix elegance is exhibited in the old dictum that the whole is worth more than the sum of the parts. Unix's elegance does, however, put an expectation on users: They must use their imaginations in defining whatever new tools they needs or desire. Unix provides simple tools that do simple things. The elegance derives in large part from what the programmer brings to the tools in terms of vision and design. The tools are there available for use. The

programmer can build a mansion or a doghouse. It is this toolbox approach that brings one writer to say, "Unix is a set of tools for smart people" (Hahn, 1994, p. 13).

Consistency

Consistency abounds in Unix. It can be seen in the pipes and filters concept, in the consistent and ubiquitous application of the file metaphor in Unix I/O. As Kernighan and Pike put it, "Instead of creating distinctions, the UNIX system tries to efface them" (Kernighan & Pike, 1984, p. 47). What this means in effect is that because files can have permissions, why can't a message queue also have permissions as well (an interprocess communication mechanism that provides a list of events that can be stored by one program and retrieved by another)? The answer is, it can. Whenever a new feature was to be added, an examination of existing patterns in the system would be performed to see if some precedent had already been set; if it was, it would be imitated.

Don't Assume A Priori That Your Users Are Idiots

Unlike some operating systems, Unix was developed from the ground up by programmers for software development. There was a tacit assumption from the very beginning that the people who would be working with Unix, installing it, programming on it, were going to be intelligent people. This was a fundamental expectation. As Dennis Ritchie once put it, "UNIX is simple and coherent, but it takes a genius (or at any rate, a programmer) to understand and appreciate its simplicity" (Vahalia, 1996, p. 16). So Unix offered power to the people, but the people were expected to ante the intellectual goods and creativity to leverage that power successfully, without shooting their foot off in the process.

Empower the User with Choice

Unix is all about choice. Unix users can choose which shell to use. They can also choose from any of a number of text editors. Or they can, like Bill Joy, simply write their own if they have the inclination. The user has a choice of a number of mail programs for reading e-mail. They can choose to read mail in the emacs editor or use other mail clients such as elm, pine, and mh.

Even the particular version of Unix is available as a choice, now that free versions are available for different platforms. For instance, a person with an Intel-based computer can choose to install Sun Solaris/X86, SCO Unix, any of a number of distributions of Linux, or a few different flavors of BSD (FreeBSD being probably the most popular). A version of Linux is available for the PowerPC, as well as for Sun's own Sparc computers. Each of these various flavors of Unix will have different capabilities, but they will all offer the essentials of Unix and demonstrate the emphasis on choice within the Unix philosophy.

CONCLUSION

In the end, what is it about the Unix operating system that so captured the hearts and minds of over a quarter century of computer programmers and users? Why do veteran hackers have fake license plates that simply say, "UNIX: Live Free or Die"? It has to do with several things, but most important among them would be freedom, imagination, and the ability to envision various futures. It has to do in the end with a penchant for trust, trusting users with the ability to build and therefore achieve remarkable things using simple tools. To offer an environment to achieve greatness, great freedom requires great responsibility, and Unix has always offered as much danger as possibility. Unix had from the beginning something Alan Kay once described (talking about the Smalltalk programming language) as "great thinking patterns and deep beauty" built in. It offered the world a chance to be creative, a trust and gift that was devoured by thousands of bright people hungering for precisely that kind of participation and community. Unix was, and still is, thanks to grassroots efforts like Linux, the FSF, and FreeBSD, all about *belonging and contributing,* and just plain fun. *Vive la beauté et la liberté.*

GLOSSARY

Application programmer interface (API) A library of functions that are callable as routines.

Bit The smallest unit of information on a digital computer, standing for binary digit. A bit can have the value 1 or 0, meaning on or off.

Boot The process of initializing the operating system on a computer, by initializing hardware and establishing a sane environment in which users can work on the system.

Berkeley Software Distribution (BSD) One of the two major variants in Unix, begun during the mid-1970s by the Computer Systems Research Group at the University of California, Berkeley.

Byte Eight bits of data on a digital computer. A byte generally contains a single character (like the character 'a') or a small integer (generally less than 128 or 256).

Command substitution The ability of a shell to execute a command and substitute the output of that command in another command.

Filter A special type of program written to take its input from standard input and write its output to standard output. Filters may be linked together using pipes. Grep and sort are two types of filters.

Function A routine that is callable in a procedural programming language, such as C. A function promotes the concept of modularity in program design.

Inode A fundamental data structure that holds information pertaining to a particular logical file on the file system. An inode holds information about the file such as its size in bytes, modification dates, security and access information, type of file, and the location of the data on the physical device, etc.

Job control The ability of the shell to run multiple programs at the same time and the facility to manage multiple programs running in the background and foreground.

Kernel The main "brain" of an operating system, providing an interface between the hardware itself and providing resource allocation services to the rest of the user system.

Multitasking The ability of an operating system to run more than one process at the same time, either on multiple physical processors or by quickly swapping tasks on and off of a single processor, giving the illusion of running multiple processes simultaneously.

Multiuser The ability of an operating system to support more than one user at a given time.

Operating system The fundamental control program on a computer, which provides a functional environment in which users can interact with the computer.

Pipe A special connector that passes through a stream of text from one end to another. Special programs known as filters can connect their various inputs and outputs to a pipe and therefore communicate over the pipe.

Process A structure that provides an execution context for a given program, providing memory resources, terminal IO support, process state, file descriptors, and so on.

Program A compiled and linked executable file that runs natively on the operating system to execute some capability on behalf of the user.

Prompt An indicator from a shell (usually in the form of '$' or '%') that it is ready and waiting for input.

Regular expression A string composed of literal characters and special symbols that can represent a potentially larger domain of possible resolutions.

Shell The primary user interface in a Unix system, allowing the user to interact with the computer by issuing consecutive commands.

Signal A brief message that one process may use to contact another process.

System call A function in the kernel's API that offers user programs the ability to request services from a running kernel.

CROSS REFERENCES

See *Client/Server Computing; Common Gateway Interface (CGI) Scripts; Linux Operating System; Open Source Development and Licensing.*

REFERENCES

Brooks, F. (1995). *The mythical man-month.* Reading, MA: Addison-Wesley.

Daley, R. C., & Dennis, J. B. (1998). Virtual memory, processes, and sharing in Multics. *Communications of the ACM, 11,* 306–312.

Digital Unix online documentation (n.d.). Retrieved October 3, 2002, from http://www.tru64unix.compaq.com/faqs/publications/pub_page/V40D_DOCS.HTM

Gilly, D. (1986). *UNIX in a nutshell: A desktop quick reference for System V & Solaris 2.0.* Sebastapol, CA: O'Reilly & Associates.

Hahn, H. (1994). *Open computing Unix unbound.* Berkeley, CA: Osborne McGraw-Hill.

Kernighan, B., & Pike, R. (1984). *The Unix programming environment.* Englewood Cliffs, NJ: Prentice Hall.

Peek, J., Todino, G., & Strang, J. (1998). *Learning the UNIX operating system* (4th ed.). Cambridge, MA: O'Reilly & Associates.

Ritchie, D., & Thompson, K. (1978). The UNIX time-sharing system. *The Bell System Technical Journal, 57.* Retrieved April 24, 2002, from http://cm.bell-labs.com/cm/cs/who/dmr/cacm.html

Ritchie, D. (1972). Unix notes from 1972. Retrieved October 3, 2002, from http://cm.bell-labs.com/cm/cs/who/dmr/notes.html

Salus, P. (1994). *A quarter century of UNIX.* Reading, MA: Addison-Wesley.

Sobell, M. (1995). *UNIX system V: A practical guide.* Redwood City, CA: Benjamin/Cummings. (Practical Guides are also available for BSD and Solaris.)

Stevens, W. R. (1990). *UNIX network programming.* Englewood Cliffs, NJ: Prentice Hall.

The Unix reference desk (n.d.). Retrieved May 6, 2002, from http://www.geek-girl.com/unix.html

Vahalia, U. (1996). *UNIX internals: The new frontiers.* Upper Saddle River, NJ: Prentice Hall.

FURTHER READING

DuBois, P. (1995). *Using csh & tcsh.* Sebastopol, CA: O'Reilly & Associates.

Free Software Foundation's GNU Software. Retrieved April 24, 2002, from http://www.gnu.org/

Gancarz, M. (1995). *The Unix philosophy.* Boston: Digital Press.

Multics History site. Retrieved April 24, 2002, from http://www.multicians.org

Lewine, D. (1991). *POSIX programmer's guide.* Sebastopol, CA: O'Reilly and Associates.

Libes, D., & Ressler, S. (1989). *Life with Unix: A guide for everyone.* Englewood Cliffs, NJ:Prentice Hall.

Rosenblatt, B. (1993). *Learning the korn shell.* Sebastopol, CA: O'Reilly & Associates.

Organick, E. (1972). *The Multics system: An examination of its structure.* Cambridge, MA: MIT Press.

Pate, S. (1996). *UNIX internals: A practical approach.* Reading. MA: Addison-Wesley Longman.

Peek, J, O'Reilly, T., et al. (1993). *Unix power tools.* Sebastopol, CA: O'Reilly & Associates.

Silberschatz, A., & Galvin, P. (1998). *Operating system concepts* (5th ed.). Reading, MA: Addison-Wesley.

Tanenbaum, A. (2001). *Modern operating systems.* Upper Saddle River, NJ: Prentice Hall.

Unix Pocket Reference List. Retrieved October 3, 2002, from http://www.utexas.edu/cc/docs/ccrl20.html

Unix programming frequently asked questions. Retrieved May 6, 2002, from http://www.erlenstar.demon.co.uk/unix/faq_toc.html

Usability Testing: An Evaluation Process for Internet Communications

Donald E. Zimmerman, *Colorado State University*
Carol A. Akerelrea, *Colorado State University*

INTRODUCTION

As Internet programs have proliferated over the last decade, "usability testing" has emerged as a buzzword for organizations claiming to have developed user-centered products. In an atmosphere of customer-focused responses, a multitude of methods called usability testing have surfaced, some of them similar processes and others different. Terminology varies widely, with some authors calling the activity "usability testing," and others referring to it as "usability engineering," or simply "usability."

This chapter presents an overview of usability testing methods, suggests when to use the methods, and lists references for further guidance. Usability testing can be used to assess users interactions with Web sites, Internet communications, software, printed publications, and hardware. The following discussion

- provides an overview of usability methodologies,
- discusses theoretical foundations for usability testing,
- reviews different usability methodologies,
- identifies ethical and legal considerations,
- notes usability's major pitfalls and how to avoid them,
- discusses verbal protocol analysis technique in detail, and
- suggests strategies for reporting usability testing results.

WHAT IS USABILITY TESTING?

Usability testing, an evaluation research methodology, should be a systematic process of assessing the design, organization, presentation, and content of proposed Internet products. A diverse series of methodologies, as discussed later, are used today. For our discussion, individuals conducting usability testing are considered usability practitioners, and we focus on Web sites to illustrate the evaluation of Internet communications. The methods discussed here can be applied to other Internet products.

Usability testing methods can be used for formative and summative evaluations as well as basic research. Formative evaluation assesses a Web site through its development and provides guidance to designers, engineers, and programmers as they continue to develop the product. Summative evaluation assesses a Web site once finished.

Although the following discussion focuses on usability testing of Internet communications and Web sites, keep in mind that usability testing can be used in diverse communications and products: testing software packages; specific kinds of software, such as accounting and inventory software; instructions and warnings; print information, such as books, booklets, brochures, and manuals; and equipment, such as medical equipment, video recorders, and cameras.

WHY TEST THE USABILITY OF INTERNET PRODUCTS?

Hard-to-use Web sites may require lengthy learning time and waste hundreds of hours of users' time, but too often the costs of bad products are hidden. Consider an e-mail program in a company with 1,000 employees. The system and e-mails reside on a server with an inbox and multiple

subdirectories. Employees can set up directories on their personal computers to back up their e-mails. What if an employee drags an e-mail from her inbox on the server to a directory on her personal computer, but the message stays in her inbox? She tries it again and yet again, but it stays in the inbox.

Clearly, the problem represents a user behavior–system design clash. The user's assumption about how the program works and how programmers designed the program to work are in conflict. Without prior experience using the particular program, users tend to fall back on their prior experience. They repeatedly try to use the program like they have used other programs, using prior mental models of how other programs work. The result is an increasing loss of time and, on the part of users, increasing frustration because the program "won't work" when viewed from their perspective.

If each employee spends just 5 minutes trying to solve this problem, with salary, benefits, and overhead averaging $50 per hour per employee, the costs can exceed $4,000. Here are the time and dollars lost:

5 minutes × 1000 employees = 5,000 minutes
5,000 minutes/60 minutes per hour = 83.3 hours
83.3 hours × $50 an hour = $4,165

Although one problem doesn't seem excessive, consider the impact on lost time and dollars if the e-mail system has 10 such problems and 500 employees spend 20 minutes a day trying to solve the problems:

500 employees × 20 minutes each day = 5,000 minutes
5,000 minutes/60 minutes per hour = 83.3 hours/day
83 hours × 5 days a week = 415 hours/week
52 weeks/year × 415 hours = 21,580 hours
21,580 hours × $50/hour = $1,079,000 per year

The time lost trying to solve problems quickly mounts and becomes extraordinary when costs are calculated for all employees. Moreover, this does not consider the impact of the lost productivity.

Usability testing can cost from a few hundred to several thousand dollars depending on the methods, audiences, and the products. Bias and Mayhew (1994) provided a framework for justifying the costs of usability testing in more than a dozen chapters outlining issues and case studies.

FOUNDATIONS OF USABILITY TESTING

The social sciences provide a range of methodological and theoretical foundations for usability testing. If followed, they can guide usability testing, minimize pitfalls, and enhance the evaluation of Web sites.

Methodological Foundations

A wide range of social science research methods have been adapted for usability testing. One of the major usability testing methodologies, verbal protocol analyses has its early foundations in cognitive psychology and problem solving research. Psychologists Ericsson and Simon (1983, 1993) developed verbal protocol analysis to investigate how people solve problems. Their works provide a detailed, critical review of the strengths and weaknesses of verbal protocol analysis. Other methodologies include questionnaires and surveys, in-depth interviewing, participant observation, and focus groups (see Table 1).

Theoretical Foundations

Specific areas of psychology provide insights into how people interact with Web sites, process information, and think about and react to the content provided through Internet products. Bailey (1996) explored human engineering acceptable performance, human characteristics, limits and differences in sensing and responding, cognitive processing performance, and memory and motivation and then provides detailed chapters on diverse aspects of

Table 1 Selected Social Science Methods and Resources for Enhancing Usability Testing and Evaluations

METHOD	RESOURCES
Overviews of social science methods	Babbie (1998)
Case studies	Stake (1995); Yin (1994)
Content analysis	Weber (1985)
Ethnography, participant observation	Bryman (2001); Fetterman (1997)
Evaluation methodologies	Rossi, Freeman, & Wirth (1979)
Experimental design	Campbell & Stanley (1963)
Focus groups	Kruger (1994); Morgan & Kruger (1998)
In-depth interviewing	Rubin & Rubin (1995); Zimmerman & Muraski (1995)
Nominal Group Technique, Delphi Technique	Delbecq, Van de Ven, & Gustafson (1975); Moore (1987); Zimmerman & Muraski (1995)
Protocol analysis	Ericsson & Simon (1984, 1993); van Someren, Barnard, & Sandberg (1994)
Surveys	Babbie (1992, 1998); Dillman (1978, 2000); Fink (1995)
Qualitative research	Crabtree & Miller (1992); Erlandson, Harris, Skipper, & Allen (1993)
Questionnaires and surveys	Babbie (2001); Dillman (1978, 2000); Fink (1995); Salant & Dillman (1994)
Validity & reliability	Campbell & Stanley (1963)
Unobtrusive measures	Webb, Campbell, Schwartz, & Sechrest (1966, 1999).

human computer interaction. Baddeley (1999) provides a succinct overview of memory, and Barsalou (1992) provides an overview of cognitive psychology. Cognitive science provides a strong theoretical foundation for building deeper understandings of how users process all forms of information. Osherson (1995) edited a multivolume compendium with individual chapters by leading researchers in cognitive science.

Carefully developed, research and evaluations can provide heuristics or guidelines for developing communications and products. For example, the National Cancer Institute's (NCI) Web site (n.d.) provides an orientation to usability testing and then provides specific research-based guidelines. The NCI organized the guidelines around categories: design process, design considerations, content/content organizations, titles/headings, page length, page layout, font/text size, reading and scanning/links, graphics, search, navigation, software/hardware, and accessibility. For each guideline, the NCI had a group of interface and Wed design researchers develop a rating scale and rate the support of the guideline based on research and evaluation. The site provides an explanation of the rating process and criteria used.

USABILITY METHODOLOGIES

Usability testing methodologies emerged in the early 1980s as companies increasingly recognized customers were having trouble with computers. The methodologies enable usability practitioners to identify problems with hardware, software, and computer content as users interact with the technology.

By early 2002, usability testing of e-commerce Web sites became exceedingly critical. Unless e-commerce sites are easy to use, consumers will not stay long enough to purchase items (Nielsen & Norman, 2000b).

Usability protocols include Web site evaluations by software engineers and site developers, usability and human factors experts, and intended users or consumers. They take many forms. For example, software engineers may conduct cognitive walkthroughs of a conceptual Web site to visualize possible discrepancies between online tasks and users' expectations of those tasks. Usability experts may engage in a structured examination of a Web site prototype, or a heuristic evaluation, to assess its usability against generally accepted standards. Software engineers, usability professionals, and users may assess a Web site prototype collaboratively in pluralistic usability walkthroughs. In addition, variations on these protocols range from simple usability inspections and site design reviews to complex distance testing with sophisticated software programs over the Internet.

In recent years, a variety of automated ways of assessing Web site design have been proposed. Among the more commonly used is click-stream analysis (Burton & Walther, 2001; Lee & Podlaseck, 2000; Murphy, Hofacker, & Bennett, 2001) and server logging programs. Usability practitioners install a software program on the server or individual computer that captures keystrokes and page visits as a user navigates the Web site. The programs provide various data, such as the more frequently used pages, the time spent on individual pages, and the least visited pages. Usability practitioners vary in their use of such programs. That said, such programs provide promise in giving an alternative, unobtrusive measure of users' interactions with Web sites.

Although usability practitioners can use many methods, the following discussion elaborates on four examples of usability techniques. They are representative protocols modeling the range of methodologies. They are card sorting, contextual inquiry, heuristic evaluation (expert review), and verbal protocol analysis (see Table 2 for a list of references related to each protocol).

Card Sorting

Web sites often fail to transfer information to consumers because developers and users have different frames of reference—they use terms differently, may not agree on the meaning of page labels, or they may organize information differently. Card sorting provides insights into how users classify, label, and organize information. Web site developers can use these insights to design, label, and structure a Web site to match users' frames of reference.

Table 2 Usability Testing Methods and Related References

METHOD	RESOURCES
Card sorting	Fuccella (1997); Fuccella & Pizzolato (1998); Koubek & Mountjoy (1991); Martin (1999a, 1999b, 1999c); McDonald, Dearholt, Papp, & Schvaneveldt (1986); Neilsen (1993, 2000)
Contextual inquiry	Beabes & Flanders (n.d.); Beyer & Holtzblatt (1995); Holtzblatt & Jones (1993)
Heuristic evaluation	IBM (2001); Nielsen (1994); Van der Geest & Spyridakis (2000)
Verbal protocol analysis	Bailey (1996); Barum (2002); Ericsson & Simon (1984, 1993); National Cancer Insitute (2001); Preece (1993); Nielsen (1993, 2000a, 2000b); Redish & Dumas (1994); Rosson & Carroll (2002); Rubin (1994); van Someren, Barnard, & Sandberg (1994)
Usability testing	Bailey (1996); National Cancer Institute (2001); Preece (1993); Nielsen (1993); Dumas & Redish (1994); Rosson & Carroll; (2002), Rubin (1994); Society for Technical Communication (2002); Usability Professionals Association (2002)

Methodology

Although experts may differ in the way they apply the specifics of card sorting methodologies, all techniques require a representative sample of Web site participants to sort a stack of 3 × 5 index cards into groups. Each card displays the name of any of the Web site's objects or pieces of information on the site, such as downloadable code, site functions and facilities, author information, content headings, or topic information. Some cards provide additional information, others do not.

Working individually or in small groups, participants organize cards in any way that is meaningful to them, creating any number of groups with any number of cards in each group. Participants may move cards around, merge card groups, or even separate card groups as they go. When finished sorting, they label each group of cards. For hierarchical Web sites, the first sort provides the lower level of the site. To obtain a higher or more general level, participants reorganize their piles into larger groups and label them with general or comprehensive headings. At the conclusion of all sorting tasks, the participants meet individually with the researcher to describe each card group, explaining why the cards were sorted into each object category.

Data Analysis and Interpretation

Card sorting data can be interpreted manually or using computerized software programs (Fuccella & Pizzolato, 1998; IBM, 1999; Martin, 1999b; National Institutes of Standards, 2002). Card sorting produces results that may be analyzed either qualitatively or quantitatively to establish the consensus among participants' classification, labeling, and organization of information. Comparing similarities in grouping schemes across ten or more participants provides adequate information to identify an emerging site hierarchy. In the latter, statistical evaluations, such as cluster analysis or similarity matrices, reveal groupings of objects. Group interviews establish consensus for labels and terminology.

The resulting data provide a basic structure of the Web site (i.e. the number of layers, the labeling of the major branches, and any replication of content within the Web site structure). In interpreting the data, the key idea is to capture the more frequently occurring structure and labeling so that it does the best job of representing participants' ways of thinking about the Web site content and organizing it.

Contextual Inquiry

Contextual inquiry, a field research technique borrowed from anthropology, identifies how computer users interact with computers as they to do their jobs. Holtzblatt and Jones (1993) used contextual inquiry in the computer industry in the early 1990s and developed a set of principles and practices to develop user-centered system design. System engineers observed users' actions as they went about their workday to understand the types of information users needed to work. In this method, engineers acquired the data to design the hardware and software to enhance users' workflow and usage patterns. Since then, others have adapted this technique as a general method for information gathering beyond systems analysis.

Methodology

Contextual inquiry is not a step-by-step process, but a set of concepts defining how user information is collected and analyzed (Holtzblatt & Jones, 1993). It requires understanding the user's work environment, establishing a collaborative relationship between user-technology and the Web site designer, and acknowledges the perspective from which the designer approaches the inquiry.

A team of designers observes users throughout a typical workday and asks them to articulate what they are doing and thinking. The designers explore any confusion or contradictions in their assumptions about the user's workflow or activities, continually validating any new information with the user. Probing questions expand the designer's understanding of the user's job. Users lead the discussion, covering what they determine is important, without direction from the designer.

Data Analysis and Interpretation

Once the designers complete all inquiries and compile their data, they begin the analyses. They review all data, interpret it, and share ideas, issues, observations, and questions. They discuss their individual and collective foci, generating their understanding and interpretations of how users will interact with the technology. Next, they reorder and group similar ideas and thoughts together and label the groupings. The labels represent the work domains and design areas for the system. The items beneath the labels comprise the developers' design ideas and users' work details. Once complete, the designers have a full picture of computer system requirements to fit users' needs.

As usability practitioners work with the data, they begin to interpret the findings and to make sense of how users work, what they do, and how they go about their daily activities. The key then becomes translating and interpreting the findings into the implications for the design of the Web site.

Heuristic Evaluations

Heuristic evaluations, sometimes called expert reviews, are systematic inspections of Web sites judged against usability design standards. The technique identifies major usability problems and produces a comprehensive site evaluation, focusing on the site's functions as they relate to users' abilities and needs.

Methodology

Because heuristic evaluations measure multiple facets of a Web site and different usability practitioners will discover different kinds of problems, two or more usability practitioners, working alone, should review a Web site and identify its strengths and weaknesses. These experts use a predetermined checklist of design principles that produce highly usable sites. Usability practitioners first navigate the Web site to obtain a sense of the site's global presentation and an understanding of how individual elements interact. Then they navigate the Web site as many times as necessary to evaluate independent elements systematically against usability principles. While they test individual elements against a checklist, any Web site feature warranting further investigation is examined closely.

Data Analysis and Interpretation

Heuristic evaluations usually produce a list of the Web site potential usability problems. The heuristic evaluation clarifies what aspects of the site design and user behavior clash and negatively impact the Web sites usability and users' experience. The list then provides Web site developers with detailed information on specific aspects of the Web site that need improvement.

Interpretation of an expert's recommendations revolves around reviewing the guideline or principle or suggestion and then figuring out how apply it to the Web site. For example, guidelines suggest minimizing scrolling on the home page. Often this requires reformatting the page and reorganizing the basic structure with additional branches.

Verbal Protocol Analysis

Verbal, or "think aloud," protocol analysis uses a set of scenarios requiring participants to carry out tasks on the Web site. As users work, they think aloud, explaining what they are doing and why. Usability practitioners observe and videotape participants, interpret the results, and report problems to Web site developers with suggestions on how to redesign the Web site to minimize the problems. The scenarios and tasks reflect the intended use of the Web site and may require participants to search for information, fill out forms, buy products online, conduct simulations, or complete other Web site functions.

Verbal protocol analysis can be conducted in a usability laboratory or in the field. Usability laboratories typically consist of waiting, observation, and test rooms. Test rooms have personal computers connected to the Internet with video cameras, microphones, and a one-way mirror for observing participants. The observation room has video mixing units, monitors, recording equipment, and computers for logging observations and monitoring participants in the adjacent test room. Usability practitioners often debate what is needed for a laboratory, and thus some usability laboratories have a complete system and others have only the most basic components—camera, microphone, tripod, and video playback units. Usability laboratories do not need all of the latest audio and video recording and mixing technology to collect useful and helpful data. In fact, usability practitioners can use a single video camera, tripod, and microphone for collecting data in the field. Other practitioners may have top-of-the-line, portable usability laboratories with multiple cameras, tripods, microphones, mixing and recording units, laptop computers, and logging software.

Methodology

Usability specialists create scenarios with tasks reflecting the intended use of the Web site. Scenarios differ depending on the research question and objectives of the usability testing. Some scenarios have participants use a particular subarea or component of the Web site, others evaluate different page design and elements, and still others have participants navigate the depth and breadth of the Web site.

Although usability testing is often complete with individuals, some usability testing may be conducting with teams of participants, and still others may be conducted using several participants completing tasks in a laboratory at once. Participants reflecting the users of the Web site are recruited to assist with the testing. See Integrating Usability Intro the Design Process for a discussion on sampling and recruiting. Usability practitioners can collect both qualitative and quantitative data of the verbal protocol analysis session. To begin, usability practitioners observe and record participants' actions on observational logging forms, paying especially close attention to the problems participants encounter and how participants try to solve the problems.

Usability practitioners also note the beginning time when a participants begins a task, as well as the beginning time when a participant first encounters a problem and the ending time, if the participant solves the problem. They also use advanced video systems using multiple cameras can capture participants' facial expressions, body movements, and screen images. Some systems video capture cards export the Web site image to a videotaping system. Other systems use software programs to capture screen and log movements, keystrokes, and navigational sequences. Participants may be asked to complete questionnaires before and after the protocol analysis. While the participant is completing the questionnaire, the usability practitioner should write a brief summary of his or her observations, the major problems encountered, and how data were collected.

Data Analysis and Interpretation

Data analysis may be either qualitative or quantitative, depending on the data collected. For qualitative data, usability practitioners review their written notes, observational logs, videotapes, and audio comments to produce a comprehensive listing of problems in site design, layout, navigation, navigational aides, site structure, content, and readability. For quantitative data, the usability practitioner summarizes the data, usually by entering the data into statistical analysis programs and then generates summary and inferential statistics, when needed. Although the list will not resolve the problems, it provides Web site developers with the information they need to improve the site.

Interpreting the findings usually requires translating the findings into strategies to solve a particular problem. Thus, it requires trying to figure out what the underlying cause of the problem might be and then considering the alternative strategies for solving the problem and making recommendations. For example, should participants report that the download time is too long for a page, then usability practitioners should determine the file size and the speed of their connection mode and recommend strategies for reducing the file size, such as reducing the visuals file sizes.

ETHICAL AND LEGAL CONSIDERATIONS

When conducting usability evaluations, usability practitioners should not develop any usability tests that would put participants at risk. Usability practitioners should ask, "Is there any way that an individual might be at physical, psychological, or social risks by being involved in the

usability testing?" If so, consider alternative approaches. If participants could be at any small risk, usability practitioners must inform them of any potential risks posed by taking part in the usability testing.

Such procedures have been formalized by the federal government. Organizations funded by the federal government have institutional review boards (IRBs) that review all research and evaluations involving human participants. Although much of their efforts focus on medical research, IRBs review social science research and usability testing evaluations.

The IRB review process usually entails the usability practitioner preparing a packet of required forms and supporting materials. The IRB application form details methodologies, participant requirements, location of the study, explanations of the known risks, IRB contact information, benefits of involvement, and confidentiality assurances. The application packet must also include a copy of the proposal, consent forms, questionnaires, research protocols, debriefing materials, investigator resume, and other required materials. The consent forms spell out any known risks. For more information on IRBs, see the National Institutes of Health (n.d.) Office of Human Subjects Research Web page.

When conducting such usability testing and evaluations, most businesses and commercial organizations ask participants to sign a nondisclosure statement indicating they will not disclose details of the confidential materials. Legal departments can provide guidance on developing non-disclosure statements.

AVOIDING USABILITY PITFALLS

Dumas and Reddish (1999) noted a trend toward more informational usability testing based on the need to do more usability testing with fewer resources and obtaining results more quickly and that usability practitioners and Web site developers no longer need to justify the methodology or results of a particular usability testing cycle. They noted its key value in diagnosing problems. That said, Web developers and novices to usability testing need to be especially cognizant of its pitfalls and lack of consistency across usability testing studies (Gray & Salzman, 1998; Holleran, 1991).

Such findings call for a systematic approach to usability testing to avoid its many pitfalls. A solid understanding of the fundamentals of social science research methodologies (Babbie, 1992, 1998), evaluation research methodologies (Rossi, Freeman, & Wright, 1979), and specifically understanding experimental and quasi experimental design (Campbell & Stanley, 1963) and the threats to validity and reliability provide a strong background for developing systematic and careful usability testing skills.

INTEGRATING USABILITY INTO THE DESIGN PROCESS

Approaching usability testing systematically as a problem-solving process offers three advantages. First, a framework guides testing and produces a systematic project design. Second, a problem-solving approach ensures articulating the evaluation objectives to minimize wasted time collecting useless data. Third, a problem-solving approach documents the process needed to review critical steps along the way. To illustrate, the following discussion focuses on conducting verbal protocol testing of a Web site for an Intranet (see Figure 1).

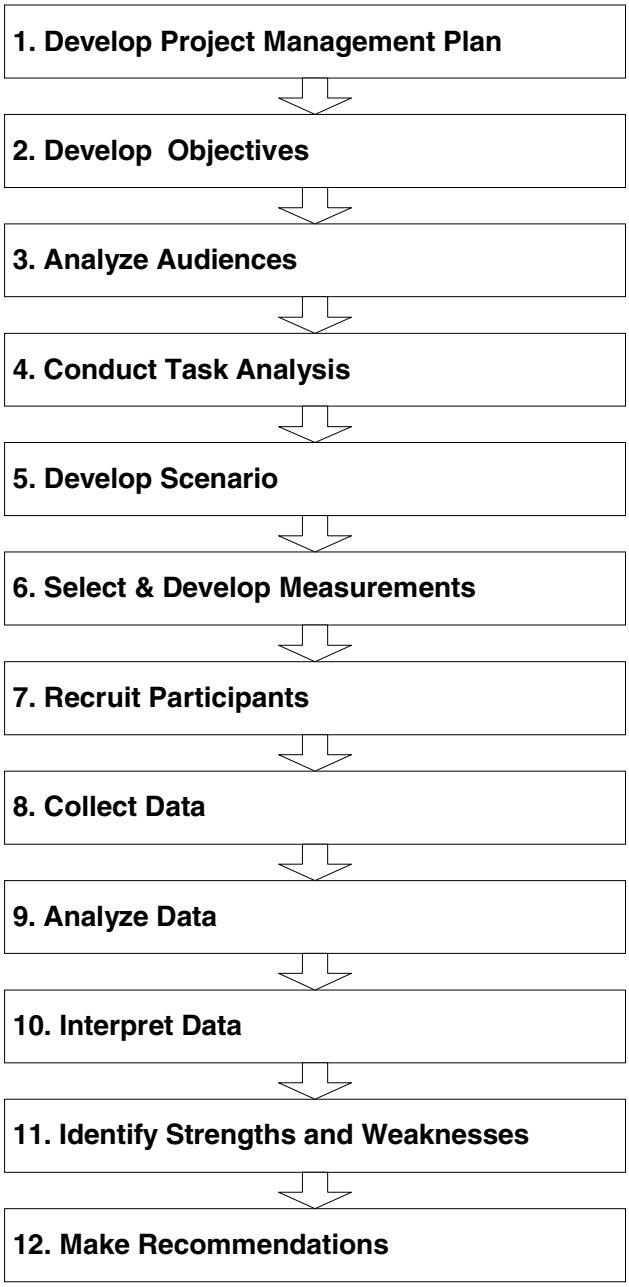


Figure 1: A systematic approach to usability testing.

Develop Project Management Plan

The key to successful usability testing is developing clearly stated objectives and a plan outlining the steps and time line for conducting the usability testing. Ideally, usability testing will be conducted throughout the conceptualization and development of the Web site, but usability testing can be conducted at any time—even for well-established

Web sites. The lessons used from conducting usability testing of existing Web sites can be used to revise the Web site and improve it. Whenever and whatever usability testing methods are used, the usability testing plan should outline the tasks, time line, and responsibilities for the usability team. Ideally, usability testing will use a team of two or more usability practitioners or assistants to help manage the logistics of collecting data.

Develop Clear Objectives

Begin by asking, "What decisions will I make based on the usability testing?" Specify the questions the site sponsor wants answered: "What do you want to know about the site? Will you collect these answers early in site development or after the site has been structured and designed? Are you testing several components or dimensions of the site, or focusing on just one? What are they?" Specific research questions might include the following: What specific problems do users encounter when trying to find information on the Web site? How severe are the problems that users encounter? How understandable do users find the information to be?

As discussed earlier, a wide range of evaluation research methodologies are available for conducting usability testing. The selection depends on the specific evaluation to be undertaken. For example, user-centered design can integrate focus groups, surveys, card sorting, protocol analyses, and follow-up surveys. Focus groups and surveys provide a better understanding of users' Web experience, content knowledge, and vocabulary. Card sorting produces an understanding of how users organize the information for the Web site and the terminology they use. Verbal protocol analyses identify user-centered problems with Web site prototypes.

Analyze Audiences

Participants should have the same characteristics as the users of the Web site. Thus, a profile is needed to ensure the recruited participants have the same background as the intended uses. Profiling the intended users, begin by asking the following questions:

- Who will be the primary users?
- Who are the secondary users?

Then ask

- What roles do audience members play in their respective organization?
- Why will they be using the Web site?
- What information might they need?
- How familiar will they be with the content terminology?
- What is the level of computer knowledge and skills of the users?
- What Internet technology skills will users have?
- What are the demographics of the users?
- When will they use the site?
- How have they accessed the Internet? By modem? T1 lines? Cable? Other systems?
- What are the users' characteristics relevant to the Web site design and content?
- What are users' usage patterns?

Although many organizations rely on informal audience analyses, others conduct formal analyses. Informal audience analyses consists of members of the Web site development team identifying the audience and then trying to describe its characteristics based on their prior experience. Formal analyses include such methods as in-depth interviews, focus groups, and surveys. Using multiple methods helps develop a better description of the intended audience and minimizes the limitations of any one methodology.

Based on in-depth ethnographic interviews, a Persona describes the flow of work (i.e., how an individual works throughout the day) and identifies his or her skills, attitudes, working environment, and goals (Cooper 1999; Goodwin, 2002). Personas guide the design of Web sites and other communications and products. Other authors have focused more narrowly on flow for modeling individuals' behaviors (Csikszentmihalyi, 1997), and others have used flow to analyze users Web behaviors of e-commerce sites (Hoffman & Novak, 1996; Novak, Hoffman, & Yung, 2000).

Conduct Task Analysis

With the objectives clearly stated, conduct a task analyses to determine the steps and problems users might encounter when using the Web site. Task analyses entails determining the sequence of steps needed to carry out specific activities on the Web site. If a potential user wants to find specific information on a Web site, identify the needed steps to interact with the Web site design and its content. Break down the tasks into their basic elements and isolate each cognitive and physical action required to complete the task. Identify potential problems users might encounter. For guidance on task analyses, see Hackos and Redish (1998), Jonassen, Hannum, and Tessmer (1989), and Zemke and Kramlinger (1987).

Develop Scenarios for Usability Testing

Next, develop scenarios to guide participants through the test. For some measurement techniques such as card sorting, the scenario can be short, simple, and straightforward. For others, such as verbal protocol analyses of a Web site, the scenario can be lengthy and detailed. Key to developing scenarios is giving participants the opportunity to navigate the Web site moving through its structure, using the different interface designs and interacting with different design elements and different content. Scenarios need to reflect different situations and need to be written in language that participants understand.

Select and Develop Measurements

After completing task analyses, select the appropriate usability testing technique. Table 3 identifies the type of information needed, at what point in Web site development the technique is often used, and then the commonly used usability testing techniques. Formative information

Table 3 Selected Methodologies for Usability Testing

TYPE OF INFORMATION NEEDED	SITE DEVELOPMENT LEVEL	USABILITY TESTS
Formative	Site objects (key parts or elements to be used in the Web site design and content) and their organization	Card sorting
Formative	Site objects, conceptual design, and their organization	Contextual inquiry
Formative	Assessment of prototype	Cognitive walkthrough
Evaluative	Assessment of prototype	Heuristic evaluation
Evaluative	Assessment of prototype	Pluralistic walkthrough
Evaluative	Assessment of prototype or finished site	Design review
Evaluative	Assessment of prototype or finished site	Verbal protocol analysis
Remote	Assessment of prototype or finished site	Automated usage tracking, and shared windowing

provides ideas for the development of the Web site development. Evaluative information provides information for revising and enhancing prototypes and finished Web site design, content, and organization. For current Web sites, evaluative information provides information for revising and redesigning Web sites to make them easier to use.

The research questions, study design, and type of usability testing suggest the variables to be measured. For example, usability practitioners often measure the time on task, the number of errors a participant makes, the number of problems encountered, and the severity of the problem. Severity can be measured in different ways, such as how successful a participant was in solving the problem or the time on task (i.e., how long it took the participant to solve the problem).

Recruit Participants

Carefully profiling Web site users and recruiting participants for usability testing is often one of the more challenging and time-consuming tasks in usability testing. In an ideal situation, having a quality list of members of the intended users, pulling a random sample, and recruiting sufficient numbers for statistical tests provides generalizable data. The lack of valid user lists and limited budgets may necessitate purposive sampling, however, which limits the number of participants and reduces the generalizability of test results.

For purposive sampling, a series of screening questions is generated to recruit participants as close to the intended user profile as possible. Researchers can recruit participants themselves, use marketing firms, or use special interest groups and local nonprofits. Marketing firms may charge $75 to $150 or more per participant, whereas special interest groups and nonprofit groups will recruit participants as fund-raising activities for $100 or more. Take care to ensure that such groups recruit participants fitting the profile of the Web site users.

As part of the recruiting process, include incentives encouraging participants to take part in the usability testing. Often termed "honorariums," the incentive can range from $10 to $100 or more depending on the intended audience of Web site and difficulty in recruiting participants. Honorariums are tokens of appreciation that compensate participants for the time required to complete the usability testing. If participants must travel to a field or laboratory site, plan to cover travel and meal costs.

Once a participant has been recruited, arrange the specific times and location of the usability testing. About a week before the usability testing session, send participants a letter confirming the time, date, and location of the usability testing. Include a map, directions, and parking instructions. A polite call the day before the usability testing session reminds participants of their session.

The number of participants depends on the methodology used. Nielsen (1993) for example, suggests discount usability testing with as few participants as possible but laces his discussion with caveats. Shneiderman (1998) labels discount usability testing as quick and dirty. Low numbers can produce erroneous information if they do not represent the intended users. If usability testing focuses on comparing different designs, different organizations of a Web site or comparing other features, recruit adequate numbers of participants to run inferential statistical tests. Rubin (1994) suggests a minimum of 10 to 12 per condition, and social scientists often recommend 15 to 20 participants for comparisons of between-subject designs. Within-subjects design can use 4 to 10 participants per group, but setting up the evaluation and running the experiment is more difficult.

For verbal protocol analyses, a practical approach suggests pretesting the methodology with three to six participants, identifying the problems common to all participants, and correcting those problems before engaging in full protocol analyses with a larger number of participants. Once the major problems are corrected, pretest again, identifying major recurring problems before conducting verbal protocol analysis with a larger number of participants.

Collect Data

Whenever possible, use multiple data collection methods to minimize the chances of drawing erroneous conclusions. Verbal protocol analyses often include direct observations of participants, videotaping their actions, recording their comments, and having them complete questionnaires. When participants arrive for testing, provide a brief overview of the evaluation, but do not make judgmental statements that may bias the test results. Brief participants on all forms, have them complete the forms, and give them the honorarium. Once they have signed the forms and completed pre-questionnaires, accompany participants to test rooms. Orient participants to the equipment and Web site they will be evaluating, give them the scenarios, and review the scenarios with them.

When videotaping or audio taping participants' comments, make sure the video cameras and audio equipment work properly and that you have sufficient tape to complete the session. Brief participants on their tasks, and remind them to think aloud as they work. As participants work, you may need to remind them to keep talking and telling you what they are thinking.

While participants are working, remain as unobtrusive as possible, yet remain available to participants if they need help or become anxious. Use observational logs (see Figure 2) to record your observations and notes of the problems they encounter. Should a participant become nervous or irritated, stop the testing. After participants have left each verbal protocol analyses session, write a succinct summary of the test session, your observations, the participant's response to testing, and the problems the participant encountered.

Analyze Data

Analyzing data can involve both qualitative and quantitative tasks. For example, textual notes from exit interviews require qualitative analysis. Qualitative analysis requires reading through the responses, familiarizing oneself with the responses, and then summarizing the findings verbally. On the other hand, questionnaires with set responses generate quantitative data. Summarizing responses to such questions requires entering the data into a spreadsheet or statistical analyses program, such as SPSS or SAS, for statistical analyses.

Use triangulation to validate findings by comparing data from different methodologies. For example, a researcher may validate data about a site's navigational aides from his or her observation notes, screen tracking software, open-ended questions, and the usability questionnaire.

Interpret the Data

Consider prioritizing the severity of the problems. Resources, staff members, and the time available to correct the problems are often limited, so many organizations must rank order the problems and correct only the more serious or critical ones or those that will affect the largest number of users. Severity ratings can be based on the time required to complete a task, the number of users who might encounter the problem, the negative impact on

Usability Observation Form **Page Number _____**

Participant No. _____

Observational Sheet _____

Date: _____

Observer: _____

Search engine _____

Task Number	Beginning Time	End Time	Problem/ how resolved

Figure 2: Observational sheet for recording participants actions when carrying out tasks.

users, and the negative impact on users' perception of the Web site. When you have analyzed all data, summarize your findings.

Identify Strengths and Weaknesses

Begin by identifying what participants liked about the Web site and what they found easy to use. Then proceed to identifying the recurring problems that many users encountered. Focus first on the recurring problems encountered when users interacted with the Web site. For each problem indicate a severity rating—how critical is it? Will it case critical problems for users? How many users might encounter the problems?

Make Recommendations

Try to develop multiple suggestions to overcome the identified problems by providing a range of possible solutions, illustrating those solutions with examples from other Web sites and providing examples of the coding used to achieve the solutions. Key to solving the problem is obtaining consensus among the Web development team. Thus, providing multiple solutions engages members of the Web development team to think about the potential solutions, and they may develop other solutions that will solve the problem. In the long run, the best solutions are ones that overcome the problem and are adopted by the Web development team.

REPORTING ON USABILITY TESTING RESULTS

Prepare a written report of the usability testing, a formal presentation including screen captures to illustrate the points, and possibly a videotape illustrating participants interacting with the Web site.

Prepare the Report

Begin by writing the technical report. It should include an executive summary, introduction, methods, findings, discussion, recommendations, and conclusions. Develop the tables and figures of the key findings, then write the findings around the tables and figures. We now describe each component of the report in brief.

Executive Summary

A one- or two-page executive summary, usually written after the report is drafted, condenses the document into a quick and easy-to-read explanation of the project and its key findings. Many readers read no further than the executive summary, so succinctly review the project's background, state the objectives, provide a brief summary of the methods, identify the problems that participants encountered, describe the site features causing the problems, and provide suggestions for eliminating the problems. Use bullet lists and boldface text to stress the most important information.

Introduction

The introduction provides a brief summary of the researching or evaluation setting, briefly describes the Web site and its purpose, identifies the intended users, and then articulates the research objectives.

Methods

The methods section describes how the usability testing was conducted by profiling the intended users, describing the sampling strategy of participants, detailing the data collection, and explaining the analysis. In this section, researchers should clearly describe all aspects of the methodology and answer the "who, what, when, where, and how" of the testing.

Findings and Discussion

Usability practitioners disagree on how to present findings. Some argue for presenting the findings and discussion separately, others for integrating the two. For an integrated approach, begin with a review of the positive characteristics of the Web site and then identify the problems participants encountered. Introduce problems in rank order, identifying the most critical or most frequently occurring problems first. Keep in mind that some problems may be critical to users' success in using the product but may not occur as frequently as other problems.

Consider the education and background of the intended readers of the report and presentation when selecting the qualitative and quantitative data. Provide a qualitative overview of the problems and then add tables and figures supporting the key points made. Keep in mind that triangulation—using observation data from a different approach—helps substantiate the findings and minimizes the chances of misinterpretation of the observations. Use anecdotal comments and observations to elaborate and illustrate the identified problems and data presented. Add screen captures illustrating specific problems when possible.

Interpret the findings by answering the following questions: "What are the data saying about the Web site? How do the results answer the questions posed in the objectives?" Look for trends in data that suggest relationships between factors. Are these relationships causal or correlational? How do you know they are causal? Are items that logically belong together, such as ratings for site readability and conciseness, showing similar results? Report triangulated data to validate the findings. If the results from methods differ, explain why. Provide interesting examples to enhance relevance of the data interpretations.

Conclusions

Identify the major strengths, highlight the major problems, and provide recommendations. Consider using bullet lists to highlight the key findings.

PREPARING THE PRESENTATION

Develop a 20- to 30-minute presentation using such presentation software as Microsoft's PowerPoint. Organize the presentation similar to the report, providing succinct data table and data presentation, interpreting the results, and then highlighting major recommendations. Use screen captures to illustrate key points and video clicks, if needed. In the presentation, leave 15 to 30 minutes to

answer questions and explore solutions to the identified problems with the Web site developers.

PRODUCING A VIDEOTAPE

In some cases, preparing a short videotape of users interacting with the Web site provides a powerful illustration of the problems participants encountered and how they struggled to solve the problems. Such short clips can reinforce the points made in the report. With advances in MS PowerPoint and other presentation software, video clips can be integrated into the PowerPoint file.

CONCLUSION

In this chapter, we have introduced basic concepts of usability testing and provided resources for readers who want to consider conducting usability testing or purchasing such services. Usability testing, if handled carefully, can provide ways to develop user-centered Web sites, software, printed publications, and hardware for diverse purposes.

GLOSSARY

Card sorting Used in the iterative Web design to identify the hierarchical organization and terminology for a group of concepts based on users' frame of reference. Users sort and group items listed on a set of cards according to how they would collect and label the concepts.

Experimental design The protocol or plan for conducting research.

Heuristic evaluation The systematic Web site review conducted by one or more usability experts to assess the site's usability according to generally accepted standards. The technique identifies major usability problems and produces a comprehensive site evaluation.

Institution review boards (IRB) Organizational committees, required by U.S. government regulations, that review research project designs to minimize participants' risk, to ensure participants' understanding of the risks of being involved in a research project, and to ensure participant's rights.

Participants Individuals who are tested or observed as part of a research project and from whom data are gathered.

Practitioners Individuals conducting usability testing. They may be software or systems engineers, human factors engineers, computer programmers, graphical user interface developers, or Web designers.

Purposive sample A nonrandom sample of participants intentionally chosen to meet a set of criteria or characteristics of a larger population.

Qualitative research The analysis of observations or events in the field that does not rely on numerical measurement and analyses of variables.

Quantitative research Research that measures variables collected during observations and experiences that focuses on the frequency and circumstances under which a variable occurs.

Random sample The selection of small group or subset of a larger population of research participants chosen by a method ensuring everyone in the population is equally likely to be selected.

Reliability The degree to which a measurement is consistent in producing the same or nearly same answer at different times.

Users Any of a group of individuals who employ a piece of technology to achieve a goal. Over time, users establish expectations related to the use of the technology.

Validity The degree to which a measurement actually assesses the concept it is believed to test.

Verbal protocol analysis A technique for investigating how people solve problems. Applied in usability testing, verbal or "think aloud" protocol uses a set of scenarios requiring participants to carry out tasks on a Web site. As they work, they verbalize their thoughts, explaining what they are doing and why.

CROSS REFERENCES

See *Electronic Commerce and Electronic Business; Human Factors and Ergonomics; Universally Accessible Web Resources: Designing for People with Disabilities; Web Search Fundamentals*.

REFERENCES

Babbie, E. (1992). *The practice of social research.* Belmont, CA: Wadsworth.

Babbie, E. (1998). *The practice of social research* (8th ed.). Belmont, CA: Wadsworth.

Babbie, E. (2001). *The practice of social research* (9th ed.). Belmont, CA: Wadsworth.

Baddeley, A. D. (1999). *Essentials of human memory.* Hove, England: Psychology Press.

Bailey, R. W. (1996). *Human engineering performance.* Upper Saddle River, NJ: Prentice Hall.

Barasalou, L. W. (1992). *Cognitive psychology: An overview for cognitive scientists.* Mahwah, NJ: Erlbaum.

Barum, C. M. (2002). *Usability testing and research.* New York: Longman.

Beabes M. S. & Flanders, A. (1995) Experiences with using contextual inquiry to design information. *Technical Communication, 42,* 409–420.

Beyer, H. R., & Holtzblatt, K. (1995, May). Apprenticing with the customer. *Communications of the ACM, 38*, 45–52.

Bias, R. G., & Mayhew, D. (1994). *Cost-justifying usability.* Boston, MA: Academic Press.

Bryman, A. E. (Ed.). (2001). *Ethnography* (4 Vols.). Thousand Oaks, CA: Sage.

Burton, M. C., & Walther, J. B. (2001). The value of web log data in use-based design and testing. *Journal of Computer Mediated Communication, 6.* Retrieved March 9, 2003, from http://www.ascusc.org/jcmc/vol6/issue3/burton.html

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research.* Chicago: Rand McNally.

Cooper, A. (1999). *The inmates are running the asylum.* Indianapolis, IN: Sams.

Crabtree, B. F., & Miller, W. L. (1992). *Doing qualitative research.* Newbury Park, CA: Sage.

Csikszentmihalyi, M. (1997). *Finding flow: The psychology of engagement with everyday life*. New York: Basic Books.

Delbecq, A. L., Van de Ven, A. H., & Gustafson, D. H.. (1975). *Group techniques for program planning.* Glenview, IL: Scott Foresman.

Dillman, D. A. (2000). *Mail and Internet surveys: The tailored design method.* New York: Wiley.

Dillman, D. A. (1978). *Mail and telephone surveys.* New York: Wiley-Interscience.

Dumas, J. S., & Redish, J. C. (1994). *A practical guide to usability.* Norwood, NJ: Ablex.

Dumas, J. S., & Redish, J. C. (1999). *A practical guide to usability* (2nd ed.). Norwood, NJ: Ablex.

Ericsson, K. A., & Simon, H. A. (1984). *Protocol analysis: Verbal reports as data.* Cambridge, MA: MIT Press.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data* (Rev. ed.). Cambridge, MA: MIT Press.

Erlandson, D. A., Harris, E. L., Skipper, B. L., & Allen, S. (1993). *Doing naturalistic inquiry.* Newbury Park, CA: Sage.

Fetterman, D. M. (1997). *Ethnography: Step by step*. Thousand Oaks, CA: Sage.

Fink, A. (1995). *The survey kit.* Thousand Oaks, CA: Sage.

Fuccella, J. (1997, October). Using user centered design methods to create and design usable Web sites. In *Proceedings of the 15th International Conference on Computer Documentation* (pp. 69–77). New York: Association of Computing Machinery.

Fuccella, J., & Pizzolato, J. (1998, June 18). Creating Web site designs based on user expectations and feedback. *Internetworking, 1*, 1–10.

Goodwin, K. (2002). *Perfecting your personas.* Retrieved August 8, 2002, from http://www.cooper.com/newsletters/2001_07/perfecting_your_personas.htm

Gray, W. D., & Salzman, M. C. (1998). Damaged merchandize? A review of experiments that compared usability evaluation methods. *Human-Computer Interaction, 12*, 203–261.

Hackos, J. T., & Redish, J. C. (1998). *User and task analysis for interface design.* New York: Wiley.

Hoffman, D. L., & Novak, T. P. (1996). Marketing in hypermedia computer-mediated environments: Conceptual foundations. *Journal of Marketing, 60,* 50–68.

Holleran, P. A. (1991). A methodological note on pitfalls in usability testing. *Behaviour & Information Technology, 10*, 345–357.

Holtzblatt, K., & Jones, S. (1993). *Contextual inquiry: A participatory technique for system design.* Hillsdale, NJ: Erlbaum.

Jonassen, D. H., Hannum, W. H., & Tessmer, M. (1989). *Handbook of task analysis procedures.* New York: Praeger.

Koubek, R., & Mountjoy, D. (1991). *Toward a model of knowledge representation and a computer analysis of knowledge measurement techniques.* Springfield, VA: National Technical Information Service.

Kruger, R. A. (1994). *Focus groups.* Thousand Oaks, CA: Sage.

Lee, J., & Podlaseck, M. (2000, January). Visualization and analysis of clickstream data of online stores for understanding web merchandising. *IBM Institute of Advanced Commerce.* Retrieved March 9, 2003, from http://www.ibm.com/iac/papers/lee.pdf

Martin, S. (1999a, June). *EZSort (beta release)*. Retrieved January 1, 2001, from http://www-3.ibm.com/ibm/easy/eou_ext.nsf/Publish/649

Martin, S. (1999b, June). *Cluster analysis for web site organization*. Retrieved January 1, 2001, from http://www-3.ibm.com/ibm/easy/eou_ext.nsf/Publish/649

Martin, S. (1999c, December). Cluster analysis for Web site organization. *Internetworking, 2*(3), 1–10. Also available at http://www.InternetTG.org/

McDonald, J. E., Dearholt, D. W., Paap, K. R., & Schvaneveldt, R. S. (1986, April 13–17). A formal interface design methodology based on user knowledge. In *CHI'86 Proceedings* (Computer Human Interaction). Boston: Association of Computing Machinery.

Moore, C. M. (1987). *Group techniques for idea building.* Thousand Oaks, CA: Sage.

Morgan, D. L., & Kruger, R. A. (1998). *The focus group kit*. Thousand Oaks, CA: Sage.

Murphy, J., Hofacker, C. F., & Bennett, M. (2001). Website generated market-research data: Tracing the tracks left behind by visitors. *Cornell Hotel and Restaurant Administrative Quarterly, 42,*82–91.

National Cancer Institute (n.d.) Improving the communication of cancer research. Retrieved August 12, 2002, from http://www.usability.gov/

National Cancer Institute (n.d.). Research-based Web and usability guidelines. Retrieved April 12, 2002, from http://www.usability.gov/guidelines/index.html

National Institutes of Health (n.d.). Retrieved May 15, 2002, from http://ohsr.or.nih.gov

Nielsen, J. (1993). *Usability engineering.* Boston: AP Professional.

Nielsen, J. (1994). Heuristic evaluation. In J. Nielsen (Ed.), *Usability inspection methods* (pp. 25–62). New York: Wiley.

Nielsen, J. (2000a). *Designing Web usability.* Indianapolis, IN: New Riders.

Nielsen, J. (2000b). Users first: How to structure your Website. *Design/usability* [online]. Retrieved December 26, 2000, from www.zdnet.com/devhead/stories/articles/0,4413,2253113,00.html

Nielsen, J., & Norman, D. (2000, February 14). Usability on the Web isn't a luxury. *Information Week.* p. 16. Retrieved March 9, 2003, http://www.informationweek.com/773/web.htm

Nielsen, J. (1993). *Usability engineering.* San Francisco, CA: Morgan Kaufmann.

National Institutes of Standards. (2002). Web metrics test bed. Department of Commerce. Retrieved August 12, 2002, from http://zing.ncsl.nist.gov/WebTools/index.html

Novak, T. P., Hoffman, D. L., & Yung, Y. F. (2000). Measuring the customer experience in online environments: A structural modeling approach. *Marketing Science, 19,* 22–42.

Osherson, D. N. (Ed). (1995). *An invitation to cognitive science.* Cambridge, MA: MIT Press.

Preece, J. (1993). *A guide to usability.* Reading, MA: Addison-Wesley.

Rossi, P. H., Freeman, H. E., & Wright, S. R. (1979). *Evaluation: A systematic approach.* Beverly Hills, CA: Sage.

Rosson, M. B., & Carroll, J. M. (2002). *Usability engineering.* San Francisco: Morgan Kaufman.

Ruben, J. (1994). *Handbook of usability testing.* New York: Wiley.

Rubin, H. J., & Rubin, I. S. (1995). *Qualitative interviewing: The art of hearing data.* Thousand Oaks, CA: Sage.

Salant, P. A., & Dillman, D. A. (1994). *How to conduct your own survey.* New York: Wiley.

Shneiderman, B. (1998). *Designing the user interface; Strategies for effective human-computer interaction.* Reading, MA: Addison-Wesley.

Stake, R. E. (1995). *The art of case study research.* Thousand Oaks, CA: Sage.

Society for Technical Communication Usability Special Interest Group (2002, 15 May). Retrieved May 15, 2002, from http://www.stcsig.org/usability/index.html

Usability Professionals Association (2002, May 15). Retrieved May 15, 2002, from http://www.upassoc.org

Van der Geest, T., & Spyridakis, J. H. (2000). An introduction to this special issue: Developing heuristics for Web communication. *Technical Communication, 47*, 301–410.

van Someren, M. W., Barnard, Y. F., & Sandberg, J. A. C. (1994). *The think aloud method.* Boston: Academic Press.

Webb, E. J., Campbell, D. T., Schwartz, R. D., & Sechrest, L. (1966). *Unobtrusive measures: Non reactive research in the social sciences.* Chicago: Rand McNally.

Webb, E. J., Campbell, D. T., Schwartz, R. D., & Sechrest, L. (1999). *Unobtrusive measures* (2nd ed.). Thousand Oaks, CA: Sage.

Weber, R. (1985). *Basic content analysis.* Thousand Oaks, CA: Sage.

Yin, R. K. (1994). *Case study research.* Thousand Oaks, CA: Sage.

Zemke, R., & Kramlinger, T. (1987). *Figuring things out.* Reading, MA: Addison-Wesley.

Zimmerman, D. E., & Muraski, M. L. (1995). *The elements of information gathering.* Phoenix, AZ: Oryx.

FURTHER READING

Bias, R. G. (1994). The pluralistic usability walkthrough: Coordinated empathies. J. Neilsen & R. L. Mack (Eds.), (pp. 63–76). New York: Wiley.

Coe, M. (1996). *Human factors for technical communicators.* New York: Wiley.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design & Analysis issues for field settings.* Boston: Houghton Mifflin.

Gleitman, L. R., & Liberman (eds). (1995). *An invitation to cognitive science: Language* (Vol.1, 2nd ed.). Cambrigde, MA: MIT Press.

Kosslyn, S. M., & Osherson, D. N. (1995). *An invitation to cognitive science: Visual cognition* (Vol. 2, 2nd ed.). Cambridge, MA: MIT Press.

Kvale, S. (1996). *Interviews: An introduction to qualitative research interviewing.* Thousand Oaks, CA: Sage.

Linstone, H. A., & Turoff, M. (. (1975). *The Delphi method: Techniques and applications.* Reading, MA: Addison-Wesley.

National Academy Press. (1997). *More than screen deep.* Washington, DC: Author.

Nielsen, J., & Mack, R. L. (1994). *Usability inspection methods.* New York: Wiley.

Smith, E. E., & Osherson, D. N. (1995). *An invitation to cognitive science: Thinking* (Vol. 3, 2nd ed.). Cambridge, MA: MIT Press.

V

Value Chain Analysis

Brad Kleindl, *Missouri Southern State University–Joplin*

INTRODUCTION

Michael Porter developed the value chain strategic framework in his 1985 book *Competitive Advantage: Creating and Sustaining Superior Performance*. Porter recognized that firms do not consist of isolated sets of functions; instead, they are chains of value-creating activities that gain competitive advantage by delivering value to their customers (Porter, 1985). These activities are tied together through a communication process that extends from a firm backward to suppliers and forward to customers. Porter's original work was influenced by the communication technology of the mid-1980s including computing, telemarketing, EDI systems, and databases.

As the 1980s moved into the Internet era of the early 1990s and then to the e-commerce era of the later 1990s, a number of new information technology (IT) tools and techniques emerged. These go beyond the communication revolution of the Internet and include new business practices that are reshaping where firms create value and how functions are linked together in business models (Chabrow, 2000). The tools and techniques outlined in this encyclopedia are becoming the weapons of choice in restructuring value chains to gain competitive advantage. These new value strategies are also fostering the restructuring of business models as businesses learn how to organize their value chain components to provide value and develop and maintain long-term relationships with customers.

Firms such as Amazon.com, Dell, eBay, and others have pioneered online retail sales models that link technology functions from a Web-based point of sale backward to inventory and suppliers and forward to after-sales service and support. Covisint has been developed and supported by the automotive industry to bring efficiencies to the supply chain. These firms have identified business models and activities that deliver value to their customers. This chapter will outline how the use of tools and techniques outlined in this encyclopedia relates to Porter's value chain framework, indicate how these activities are linked in business models, and explain how these concepts relate to an e-commerce value chain.

BASICS OF THE VALUE CHAIN FRAMEWORK

In *Competitive Advantage* Porter analyzes how firms gain competitive advantage. A firm's profitability is determined by the industry structure in which it operates and the sustainable competitive advantages a firm can obtain. In the value chain analysis process, Porter disaggregates the activities of buyers, suppliers, and the firm into a chain of interrelated value-creating activities. A firm's value chain is seen as consisting of nine generic interlinked activities including the *primary activities* of inbound logistics, operations, outbound logistics, marketing and sales, and service, and the *support activities* of procurement, technology development (including R&D), human resource management, and infrastructure (systems for planning, finance, quality control, information management, etc.).

Each of these broad activities of a business, such as manufacturing, marketing, and management, is subdivided into a set of interrelated activities. Each industry is likely to have its own unique value chain structure, and every firm within an industry may decide to structure its value chain differently to gain distinctive advantages. For example, firms may decide to outsource activities where they cannot provide value efficiently.

Firms can use the value chain framework to gain competitive advantages that come from performing value activities at a lower cost or to differentiate the value chain in a way that allows higher prices. A firm can evaluate competitors' value chains to determine relative strengths and weaknesses. It can also look for ways to strengthen linkages within its current value chain. Porter recognized

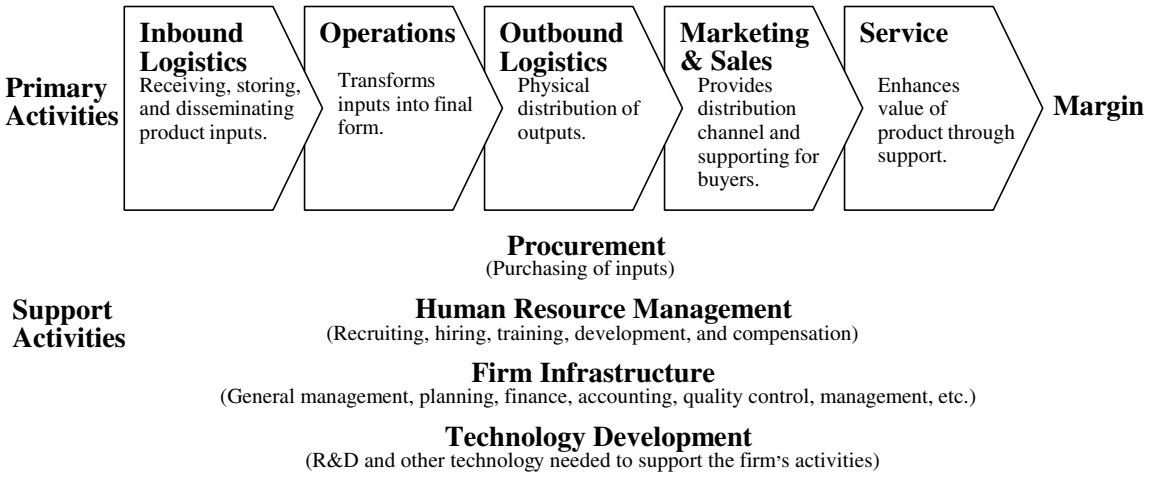


Figure 1: Porter's generic value chain for a firm.

that communication is the key to developing linkages to enhance value within the broader spectrum of suppliers', firms', and customers' value chain activities. A single order from marketing could create information for manufacturing, procurement, accounting, and service. Information could be passed backward to suppliers and forward to customers to inform them of expected shipping time.

Figure 1 outlines Porter's generic value chain. This includes a breakdown of primary activities into inbound logistics, operations, outbound logistics, marketing and sales, and service, supported by procurement, technology development, the firm's infrastructure, and human resource management. The value chain should deliver a margin, or a difference between the potential value delivered and the cost of creating that value.

In the mid-eighties firms could use information technology to enhance the value chain and gain competitive advantages. The early information revolution influenced the competitive environment in three ways: it changed the industry structure altering the rules of competition, it created competitive advantages by giving companies a means of outperforming competitors, and it spawned new businesses. Improvements in information technology would lead to increased power of buyers and increased rivalry within industries (Porter and Millar, 1985). The Internet and associated technologies such as wireless communication has increased the pace of the communication revolution. Porter (2001) added to his strategy and value chain frameworks in light of the growth of the Internet and other technology based applications. The growth of information technology has opened a new window on the value creation process.

ADDITIONAL PERSPECTIVES ON THE VALUE CHAIN CONCEPT

Cartwright and Oliver (2000) proposed a value web for firms whose products have high information content, service firms, and firms that operated in the electronic marketplace. This perspective considers value creation as taking place when several organizations share common technologies or intellectual capital to serve markets. A value web links the horizontal and vertical relationships that exist in competitive systems and recognizes that alliances can be virtual and arrangements temporary.

Rayport and Soviokla (1995) increased the importance of information within a value chain by describing a virtual value chain. In this view, a firm captures information along its entire value chain and then uses that information to enhance value and improve customer relationships. The firm treats information, such as database information, as a product that can be sold or used to enhance value.

E-COMMERCE VALUE CHAIN

This encyclopedia outlines a number of e-commerce tools and techniques that offer a new perspective to the value chain. The e-commerce value chain views information technology as part of a business's overall value chain that adds to the competitive advantages of a business (Carr, 2001; Fingar & Aronica, 2001; Rayport & Soviokla, 1995). Porter (2001) has noted that the Internet affects the value creation process in a number of ways. First, the Internet has made increased operational efficiency possible. Internet protocol (IP) standards allow flexible and unified technology systems that can be used across applications and by a firm's multiple constituencies. IP standards link the supply chain (through extranets), foster internal communication (through intranets), and communicate with external constituencies (through the Internet). One study found that on average 61% of U.S. organizations used some Internet-based business solution. This has resulted in a cumulative cost savings from 1998 to 2001 of $155 billion and is expected to produce an additional $373 billion in cost savings by 2005. In Porter's view, operational efficiency does not necessarily allow for the development of competitive advantages. In fact, when all firms use the same technology to obtain the same efficiencies, they are forced to compete on price, lowering the overall profitability in the industry.

The assessment of a firm's e-commerce value chain requires an analysis of the competitive forces within an e-commerce environment, specification of the functional business model to be used (how functions are interlinked), and the identification of the value activities that will allow the e-commerce value chain to be structured for a competitive advantage. This requires an understanding of

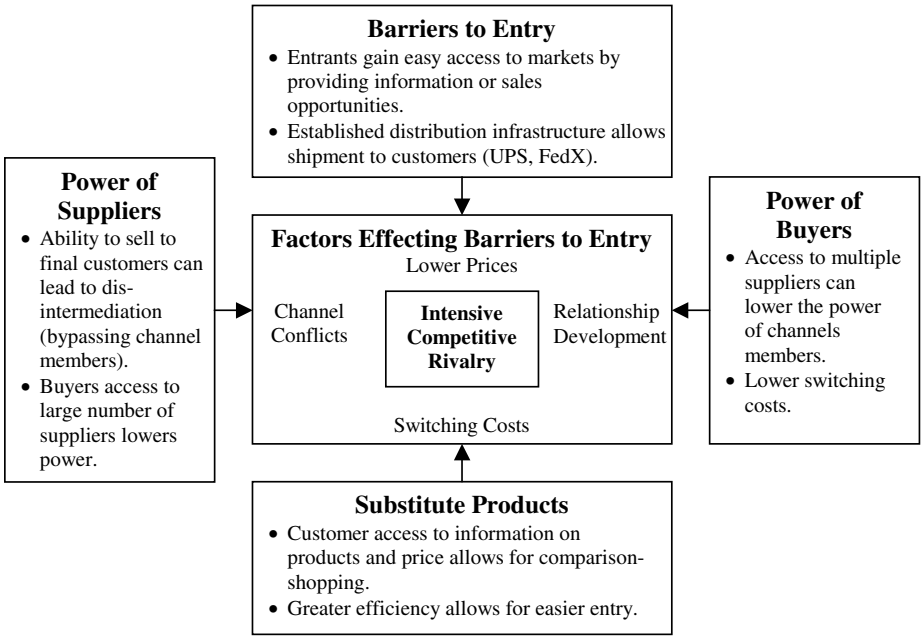


Figure 2: Information-technology-based competitive force analysis.

how to integrate information technologies within a firm's business model. Competitive advantages come not only from Porter's list of primary activities, but also from the supporting activities of the firm. New business practices require that the value chain disaggregation process identify new IT-based activities that add value to the firm. The assessment of an e-commerce value chain is relevant not only for pure-play Internet companies, but also for brick-and-mortar-based businesses.

Five-Forces Concept

Many industries operate in rapidly evolving and highly competitive environments that are forcing change in business models, business strategies, and in the distinctive competencies needed to compete. An analysis of an industry's competitive structure must be undertaken before assessing where a firm can gain sustainable competitive advantages. Drivers leading to environmental turbulence include the following:

Technological change: Computing power has increased while costs have decreased, allowing technology to be applied across a broader spectrum of products and uses.

Changing customers: The Internet and online purchasing are growing around the globe. Technology is an enabler, allowing individuals to accomplish more. Access to information has increased customers' negotiating power.

Shorter product life cycles: Rapid development of new technology, aggressive marketing, and buyers' willingness to try new products have increased new product development.

Increasing number of competitors: The distance between competitors is vanishing. Online access has increased the intensity of competition by allowing competitors into new markets.

A need for speed: Instant connectivity is becoming the norm in business-to-business applications as well as in the way consumers shop (Chabrow, 2000; Kleindl, 2003; Rosenau, 1990).

Porter (1985) used a five-forces concept to evaluate the competitive environment of an industry. Information technology has reshaped industry structure. Figure 2 illustrates how these forces interact within an e-commerce competitive framework (Kleindl, 2000; Porter, 2001). Firms face intensified competitive rivalry due to a number of factors. Information technology has shifted power for suppliers and buyers. Buyer power has increased because of easy access to competitive information and competitive products. Access to price information allows customers to negotiate leading to dynamic pricing, or the lack of fixed prices. Buyers' power has also increased because of the requirements that suppliers link through extranets where they often must engage in competitive bidding through online auction systems to win an order. Supplies have found that they can increase margins by using technology to bypass intermediaries. This has lead to channel conflicts.

Threat of entry into markets has increased because transaction costs have dropped. Low-cost and standardized IP platforms have lowered barriers to entry and switching costs for buyers. Potential entrants can use the Internet to provide information or sales opportunities and then use established distribution systems to deliver products to buyers. Buyers are able to use the Internet to find and price substitute products. The increase in the number of suppliers can lead to commoditization of markets with resulting lowering of prices.

Porter (2001) views a migration to price-based competition as one of the major outcomes of this changing industry structure. Firms can protect themselves by gaining efficiency and lowering prices, increasing switching costs, limiting the chances for channel conflicts, and

focusing on relationship development with customers. To maintain pricing power firms must differentiate and find a means of creating value for customers by evaluating the e-commerce value chain and business models.

BUSINESS MODELS

A value chain disaggregates business functions into a set of activities. A business model differs from a value chain in that it illustrates the process flows between the various value-creating activities of a firm. As with value chains, businesses must modify their models in response to change in the competitive environment. Of the variety of business modeling perspectives, a functional business model identifies the interaction of functions within a business. This modeling process allows a systematic approach to viewing the complex relationships that are required to make a business operate and can aid in identifying how the value chain components fit together.

Figure 3 illustrates a generic e-commerce model that takes advantage of the technologies outlined in this encyclopedia. Hypermedia such as Web sites can both provide and collect information between the company and the customer. Information collected in databases and mined allows the customer to be served as a market of one. Competitive pricing information allows dynamic pricing, or a lack of fixed prices. Auctions and negotiations allow dynamic pricing, but so do database systems. Amazon.com used its databases to charge different prices to different customers based on what the database specified customers would be willing to pay. This resulted in higher prices for some customers and a strong consumer backlash. Credit card companies and online billing and invoicing systems facilitate online payment flows. Customized products can be delivered through independent shippers directly to the customer. Both of these functions are often outsourced. Inventory and working capital are minimized because suppliers are linked to production needs. Leaders who understand and can leverage technology must manage these business models. The firm must also have the human capital resources necessary to operate in an e-commerce environment.

Both online-only and brick-and-mortar firms have adopted components of this e-commerce model. The generic e-commerce business model illustrates what is necessary for a firm to undertake its operations, but does not indicate areas of competitive advantage. An analysis of the e-commerce value chain allows assessment of a firm's value activities, the identification of components of the business model that can be outsourced, and a determination of where competitive advantages lie. Because margins for a business are dependent upon the customer value added, firms have an incentive to develop and maintain long-term relationships with high-lifetime-value customers. The interlinking of information on transactions and customers allows online databases to determination the value of customers and the factors that can lead to extended relationships.

Figure 4 illustrates the e-commerce value chain. A firm can gain cost advantages across a number of different areas such as lower inventory cost through JIT-based extranets, built-to-order systems, lower promotion costs through Web sites, and outsourced distribution. These cost advantages may only give firms industry parity or short-term advantages, not long-term differential advantages. The use of technology and customization from databases can allow the development and maintenance of stronger customer relationships.

A key to implementing e-commerce technology is having the proper support activities, including a management

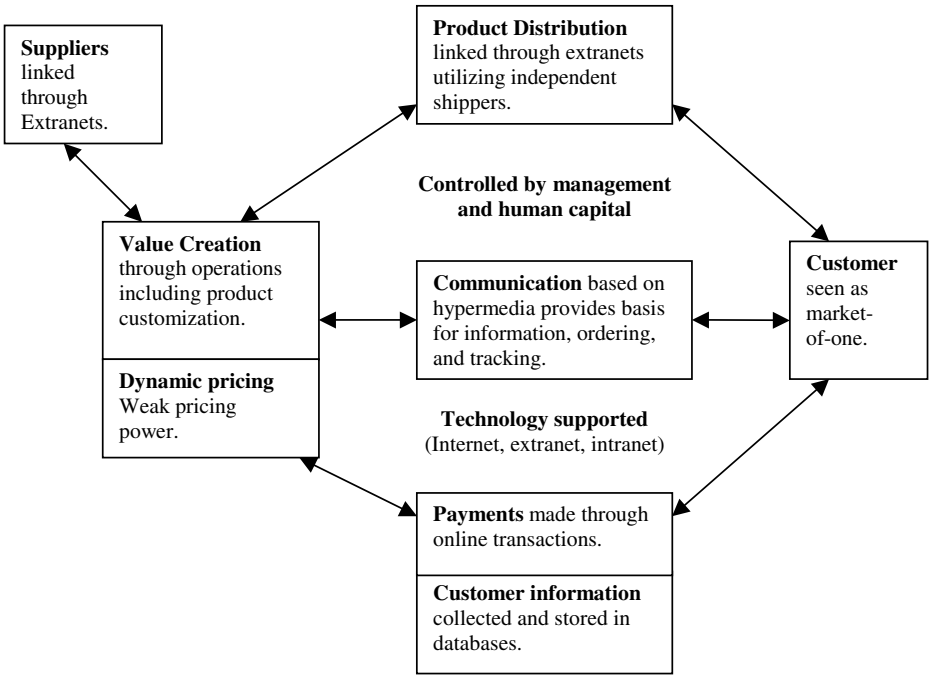


Figure 3: Generic e-commerce model.

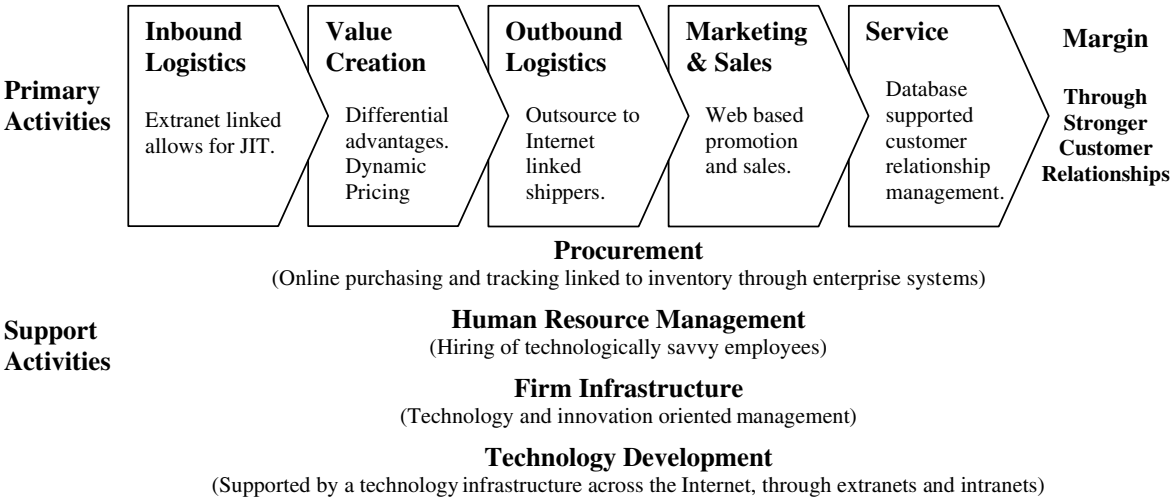


Figure 4: The e-commerce value chain.

team and employees who are willing and able to restructure business models and capitalize on the advantages found in customer databases, online access to information between buyers and sellers, rapid responses to environmental change, and proactive innovation (Downes & Mui, 1998; Slater, 1998–1999). Not all the components of the e-commerce value chain must come from within the organization itself; many e-commerce firms outsource key components of their value chains or form alliances with other businesses.

Each of the components of the generic e-commerce value chain can be disaggregated into subcomponents. Figures 5 and 6 illustrate this process by identifying the distinct marketing and sales and management/human resources sections of the value chain.

Information technology is used as a tool to enhance the e-commerce value chain and is aiding in the development of closer relationships with customers by speeding up ordering and delivery of products, improving customer service, and lowering costs (Caldwell, 1999; Szygenda, 1999). Strong customer relationships that allow a margin should be the result of e-commerce value chains.

E-COMMERCE VALUE CHAIN STRATEGIES

A value chain analysis integrates the two traditional paths to competitive advantage. Value chain analysis can help determine how a firm can obtain a low-cost (and therefore low-price) position though the gaining of efficiencies or the outsourcing of functions. Value chain analysis can also allow the determination of a differentiation strategy by identifying a unique value proposition against competitors. As has been stated, being the low-cost producer may not be enough to gain a long-term advantage, because in an electronic environment, competitors may be able to gain the same efficiencies. E-commerce technology can be both a blessing and a curse for businesses. A frictionless market implies that customers have almost perfect information and can compare prices around the world. This process can be enhanced through the use of intelligence agents to search out best prices. This forces businesses selling over the Internet and those that compete against Internet sales to lower prices or differentiate (Kuttner, 1998; Porter, 2001). Table 1 outlines a number

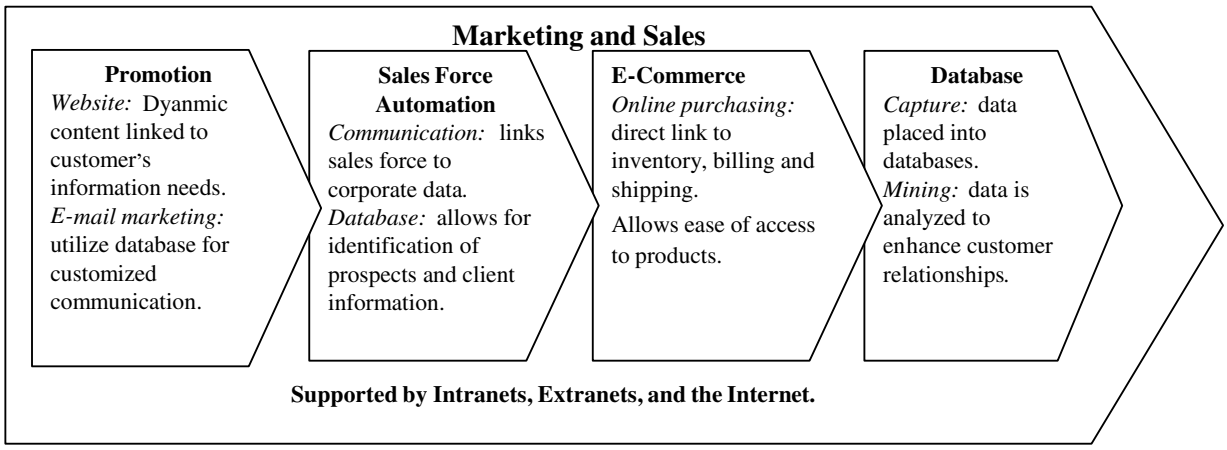


Figure 5: Marketing and sales components of the e-commerce value chain.

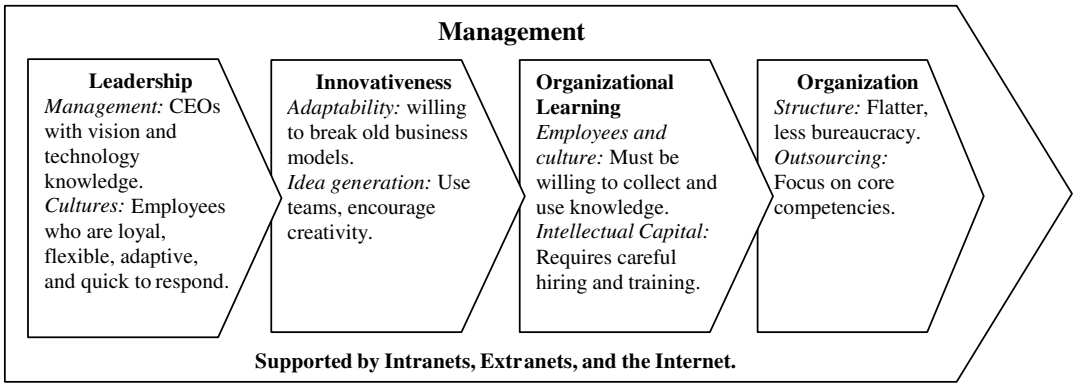


Figure 6: Management components of the e-commerce value chain.

of strategies e-commerce-based businesses are currently using to provide value and obtain strong customer relationships (Kleindl, 2003).

Firms engaged in or planning to engage in e-commerce must determine how to develop value chains to gain advantages. For e-commerce firms, speed and flexibility allow quick response to environmental change, size permits economies of scale, brand names give assurance to the buyer, and close customer relationships entice customers to return to a site. These do not guarantee long-term advantage because severe price competition may hurt all but the most efficient businesses or those with differentiated niches. Table 2 outlines the impact of a value chain analysis across different types of retail businesses (Kleindl, 2003).

VALUE CHAIN DEVELOPMENT

Identification of competitive advantages in an e-commerce marketplace requires an analysis of a firm's business model and value chain. This section outlines the process of modeling business functions and the disaggregation of functions into value chains. Developing e-commerce models requires identifying the functional process flows of a firm and then modeling how the application of e-commerce procedures can result in competitive advantage. This requires a ten-step process:

Identify the functional areas and major players.

Indicate linkage of the functional areas and the directions of the flow process.

Determine how to apply e-commerce tools and techniques to the business model.

Develop a new e-commerce model flow.

Disaggregate the business model into a value chain.

Evaluate the competitive advantages of the model by using a value chain analysis.

Analyze the firm's own value chain including the costs related to every activity.

Table 1 Methods of Differentiation

Differentiation Strategy	Value Chain Impact	
	Advantages	**Disadvantages**
Gain Speed & First Mover Advantages	This provides a number of first mover advantages including costs, meeting needs, lowering risk, and lower prices.	Firms need to be flexible to be fast. Being first increases risks and may require large amounts of capital to maintain advantages.
Build Brand Name	Gives buyers assurance when interacting with a site. Allows easy name recognition.	Requires a large amount of capital to obtain and maintain a brand name.
Portal and Market Place Development	Allows economies of scale and builds barriers to entry.	Requires large amount of capital, pushing off profitability.
Pursue Niche Strategies	Very good strategy for smaller or weaker businesses. Allows a business to focus and become an expert in one competitive arena.	Having only one niche can be risky because all of the "eggs" could be in one basket or with one customer.
Enhanced Customer Relationships	Allows businesses to build barriers to entry. By staying close to customers businesses can meet needs better.	Could result in a loss of power by the business supplying the product or service.

Table 2 Retailer Value Chain Analysis

Retailer	Value Chain Analysis
New online-only	New online-only businesses must develop their entire value chain systems from scratch.
Established online-only	Established online-only businesses such as Amazon.com may have these components developed. They may have advantages in supply chains, images as online-only firms, targeting of audiences, shipping systems for distribution, data collection and knowledge management, and managerial expertise in e-commerce.
Catalog businesses	Existing catalog businesses have these components developed. They may have advantages in supply chains, images as non-brick-and-mortar, targeting of audiences, shipping systems for distribution, knowledge management, and managerial expertise in direct marketing.
Retail chains	Existing traditional national retail chain businesses may have these components developed. National chains may have advantages in supply chains, image, targeted audiences, prime locations for distribution, information capture and knowledge management systems, and managerial expertise in retailing.
SME retailer	Existing small to medium-sized retailers may have none of these components developed for conducting business outside of the targeted market area.

Analyze the customer's value chains to determine how to fit into their value chain and determine how to develop long-term relationships.

Identify cost advantages against competitors.

Determine the likelihood of acceptance of the new model.

INDUSTRY EXAMPLES OF VIRTUAL VALUE CHAINS

Two examples of e-commerce value chains will be given. The first is an evaluation of how Dell competes against a traditionally brick-and-mortar manufacturing and retailing industry. The second is an online-only e-marketplace environment.

PC Industry

An example of a traditionally brick-and-mortar manufacturing and retailing industry model changing to an e-commerce-based model and value chain can be seen in the PC sales industry. Figure 7 illustrates the functional business model for the PC industry prior to 1998. This model relied upon traditional manufacturing, shipping, and retail sales methods.

This model followed a generic manufacturing and distribution model. Sales projections set inventory levels while continuing lowering of inventory costs made stored inventory expensive. Production was to fit retail orders, leading to a lag between customer demand and the manufacturing process. Excess inventory at the retail level would also lose value while it was stored in the retail

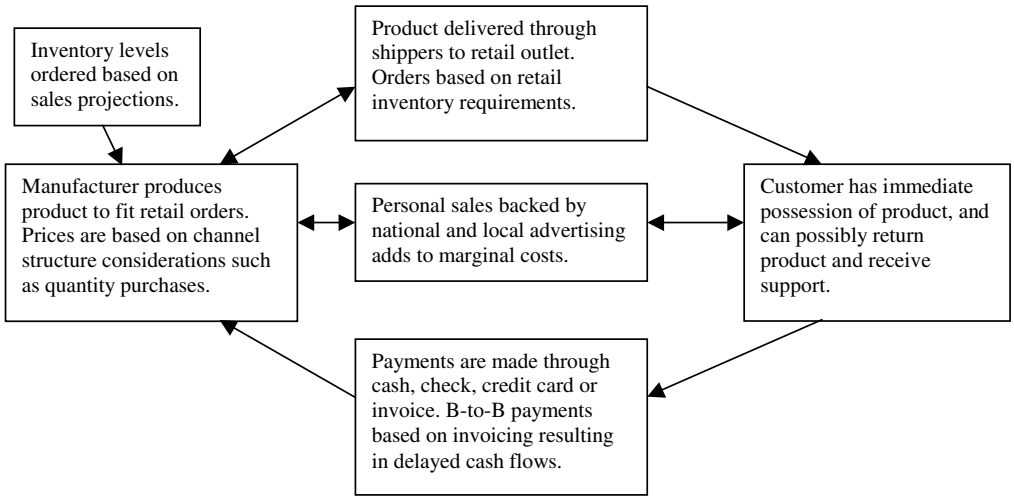


Figure 7: Pre-1998 personal computer sales functional business model. From *Strategic Electronic Marketing, Managing E-Business,* 2nd edition, by Kleindl. © 2003. Reprinted with permission of South-Western, a division of Thompson Learning: www.thompsonrights.com. Fax 800 730-2215.

outlet. National and local advertising backed by personal sales increased selling costs. The customer paid at the point of sale or through an invoice and was often able to receive immediate access to the product.

From its inception, Dell used a direct sales business model that allowed it to bypass the traditional brick-and-mortar sales model. In 1996 Dell Computers began transforming its telephone direct sales business model to an e-commerce business model. Dell evaluated its value chain and determined that its advantages lay with buying the best products from suppliers, assembling computers, and developing close relationships with customers (Magretta, 1998). Dell was able to use Internet-based technology to leverage its existing advantages (Chabrow, 2000). By 1998 online sales had reached more than $5 million a day. At the end of 2001, Dell had a 40% share of the PC industry in the United States with close to 100% of its sales transacted online (Serwer, 2002). Dell's online model utilized a number of technologies and strategies to lower costs, increase efficiencies, and enhance customer relationships. Figure 8 illustrates Dell's business model.

Dell provides value throughout its value chain. Dell's inbound logistical system requires suppliers to use extranets to link to Dell. To speed delivery Dell demands that its suppliers locate inventory within 15 minutes of the Dell factory. Dell minimizes its finished goods inventory and is therefore able to use the latest products and take advantage of dropping inventory costs.

Dell creates value through the customization of its computers to individual customers. Dell preloads required software so that its computers can be unpacked and run. As market conditions change, Dell can respond by changing its pricing.

Much of Dell's outbound logistics is outsourced. Dell outsources warehousing to third parties that specialize in running supply chains. Dell has noncustomized products such as monitors stored and delivered by shippers such as UPS. All of the companies are linked electronically to speed information flow.

Marketing and sales provide value. Most computer buyers search for information before they buy. Dell's Web site allows those buyers to move immediately to areas that interest them. In 1997, Dell received one Web visit for every phone call inquiry; by 1998 there were 3.5 Web visits for every phone inquiry. Potential buyers visit the Web site 5 to 10 times to obtain information, have their questions answered, and determine prices before they buy. Because the Web visit is considerably less expensive, the cost savings are given back to the buyer (Chadderdon, 1998). Dell also gathers customer information to determine trends in the marketplace and uses databases to support customer relationship management systems. Both individual and business buyers can receive customized interfaces.

Dell squeezed time out of every step in the business process. Dell's average sale turns into cash in less than 24 hours. By 1997, Dell was able to receive orders at 9 a.m. on one day, build the computer, and deliver it by 9 a.m. the next day. This increase in speed allowed Dell to lower inventory costs and prices to its customers. A key to this strategy was using the Internet to link Dell to both its customers and its suppliers. Developing a more efficient system resulted in operating costs at 10% of revenue compared to Hewlett Packard at 21, Gateway at 25, and Cisco at 46% (Jones, 2003). Dell was able to return $1.54 in profits to shareholders for every new dollar of capital investment in 1997, whereas Compaq returned only 59 cents and IBM 47 cents. Dell has added to its own value chain by developing a strong e-commerce-based business model. Dell Computer identified areas of key competitive advantage and changed its business model to meet its customers' desire for speed of delivery, customized products, and low prices. Dell has named this system "the model" and plans to expand into new markets including storage, networking, and information services (Jones, 2003).

Corporate customers are able to order directly from Dell using the Internet. Companies such as Pillsbury and

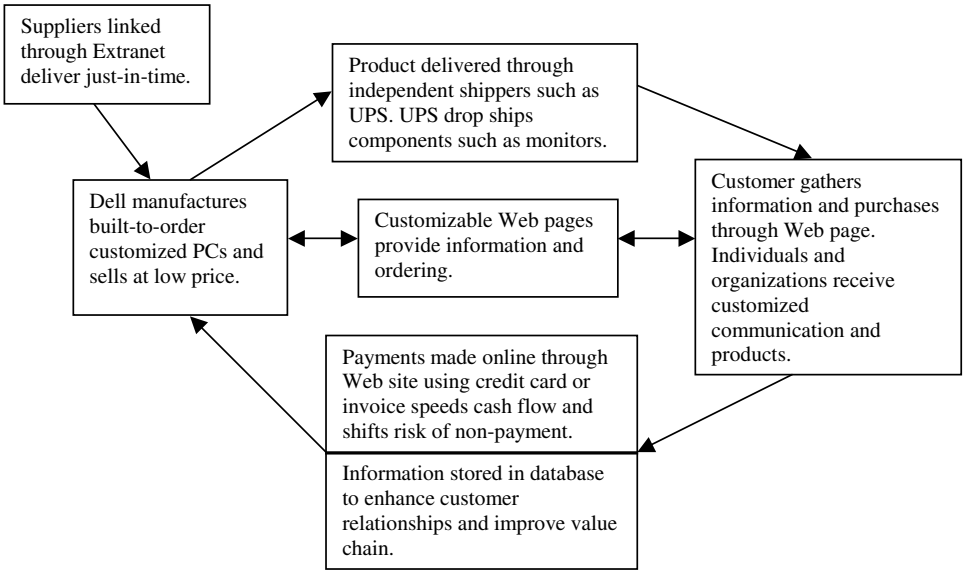


Figure 8: Dell's business model.

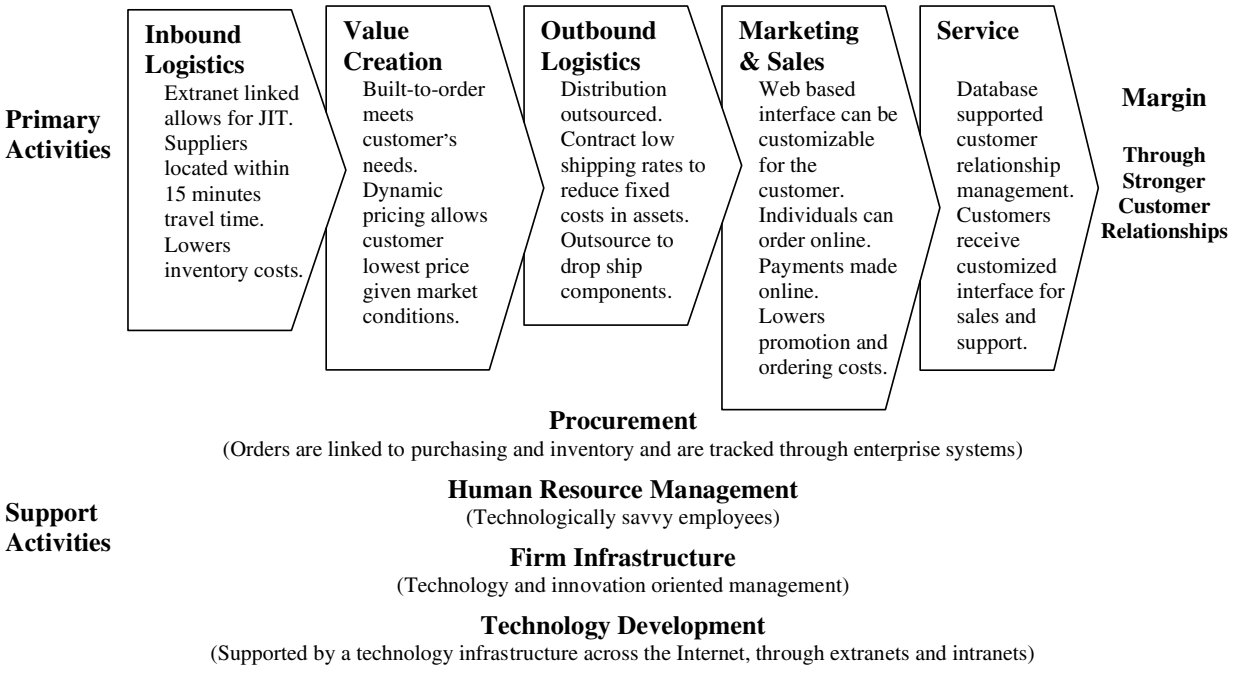


Figure 9: Dell's e-commerce value chain.

Ford have designed their own internal Web pages that link to Dell's computer system to allow online purchasing. This customization has saved corporate customers millions of dollars. Dell has moved its business model overseas to allow rapid delivery of computers in foreign markets. By 1997, 31% of Dell sales came from outside the United States.

Dell's value chain (shown in Figure 9) is designed both to be efficient (resulting in cost savings) and to gain an advantage through strong customer relationships. A key to Dell's success lies in logistical efficiencies that allow lower costs, the production of high-value customized PCs, and strong customer support.

Dell's major competitors, such as Compaq and IBM, built value chains that relied on manufacturing of components and using resellers to sell their products. Switching to the Internet for e-commerce has the potential of alienating reseller support because of channel conflicts. Each of these firms was forced to match Dell's online business model.

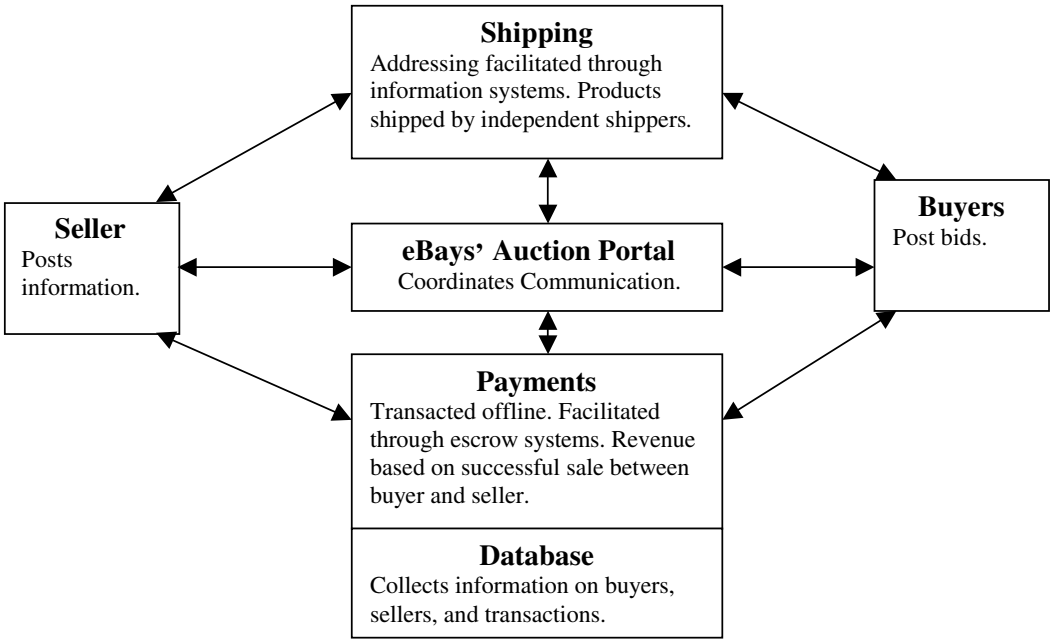


Figure 10: eBay's business model.

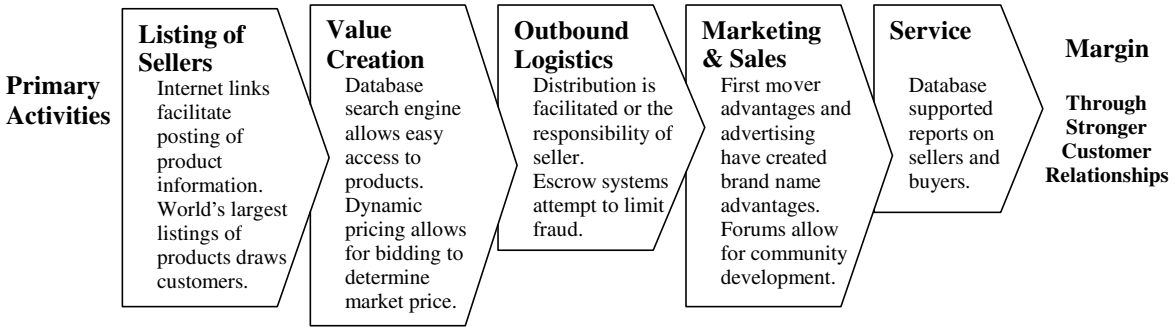


Figure 11: eBay's value chain.

E-marketplaces

Online marketplaces offer a centralized location for the trading of information, goods, services, or other commodities. E-marketplaces serve both consumer and business markets. These do not handle any physical product, instead acting only as facilitators of exchange between parties. The largest of these is eBay. In 1995 Pierre Omidyar designed a Web site called Web Auction to allow his girlfriend to sell Pez dispensers. Demand for the service was so strong that Omidyar began charging sellers. By 1998 eBay was valued at $7 billion dollars and by 1999 eBay accounted for the largest volume of consumer sales on the Internet. Figure 10 outlines eBay's business model.

In this model, the seller posts information on the product to be sold. Buyers then post bids, and after a given time the winning bidder is notified. Payments and shipping can be handled outside of eBay's system, or they may be facilitated by eBay. eBay also maintains databases on sellers and buyers in an attempt to limit fraud.

An analysis of eBay's value chain in Figure 11 gives insight into how to create value in e-marketplace business models.

Whereas Dell's business model competes against traditional business models, eBay was able to pioneer a new business model. As with many e-marketplaces, eBay's major competitive advantage lies in having the large listing of items for sale combined with the brand name recognition gained from being a first mover in the online consumer auction industry.

Customers find value by finding the products they wish to purchase. This is enhanced through the use of database search systems. Pricing is set dynamically through the auction system. Although outbound logistics is often settled outside of the e-marketplace, eBay allows escrow systems in an attempt to limit fraud.

eBay has developed a strong brand name. The use of forums and online communities helps to maintain brand name advantages. eBay also uses its database to develop reports on sellers and buyers.

SMALL AND MEDIUM-SIZED ENTERPRISES (SMES)

Small and medium-sized enterprises (SMEs) differ from large firms in that they do not have the capital and human resources of their larger competitors. To compete, SMEs have traditionally used niche strategies to gain distinctive advantages over their larger competitors. A niche strategy requires that the SME find a competitive arena

Table 3 Value Strategies for SMEs

Value Strategy	IT Manager Responses	Advantages
Improve customer service	Over 80%	Online systems allow channel members and consumers to gain access to product and inventory information.
Electronic commerce	Over 60%	This allows SMEs access to larger markets without the cost of setting up new distribution systems. It also allows the SME to target narrow markets faster then larger competitors. Lower overhead costs can be carried over to lower prices to customers.
Customer-relationship management applications	Over 50%	Online connections between the SME and its customers increase the speed of response and allow for close to instant communication. Linked extranets allow SMEs to act as virtual partners with other businesses.
Increase business-to-commerce connections (extranets)	Over 40%	SMEs can act as e-commerce intermediaries linking larger businesses with very small suppliers. Online access to inventory and supplies helps control costs.

in which larger businesses, with their greater resources, are not competing. SMEs are currently using information technology to gain distinctive advantages. Companies with 500 or fewer employees spent over $200 billion on technology products and services in 1998, more than five times as many dollars as larger companies (Caldwell & Wilde, 1998; Wilde, 1998). Small businesses are using Web sites, intranets, and e-mail at close to the same percentages as larger businesses (Wilder, 1999). E-commerce applications have been slower to be accepted; this could be due to the relatively high cost of setting up and maintaining e-commerce applications. Outsourcing e-commerce to other e-commerce companies can lower these costs (De Soto, 1998).

Table 3 outlines the results of a survey of more than 400 information technology managers worldwide and shows the main strategies that SMEs should use to compete in a global arena. In addition to these competitive strategies, smaller businesses are often more innovative, faster to respond to environmental demands, and willing to change business models in order to gain competitive advantages. The recommendations given to SMEs center on their ability to focus on the customer. SMEs do have an opportunity to hold their current customers if they can leverage e-commerce tools and techniques before their larger competitors enter their market.

CONCLUSION

E-commerce tools and techniques have become a vital part of both online-only and brick-and-mortar firms' business models. This requires that these firms consider their e-commerce value chains when determining how to gain competitive advantage in the marketplace. The environmental turbulence that firms faced in the 1990s is likely to increase as newer, faster, and more integrated information technologies such as wireless Internet are adopted by both consumers and businesses.

Gaining and maintaining competitive advantages will require that businesses constantly reevaluate their business models and their value chains. This chapter lays the foundation for this process.

GLOSSARY

Business model The basic process flow indicating how a business operates.

Commoditization of markets A situation where products are no longer seen as by buyers differentiated, resulting in a lowering of prices.

Disaggregation The process of breaking out functions in the value chain.

Disintermediation The removal of intermediaries from the distribution process.

Drop shipping The storage and shipping of products by an intermediary such as a shipper or warehouse.

Dynamic pricing The lack of fixed prices. Two examples include database systems, which can offer different prices depending on variables such as demand, buyer type, and sales channel, and negotiated pricing through auctions or other interactions.

E-commerce value chain A value chain framework where information technology is seen as a part of a business's overall value chain and adds to the competitive advantages of a business.

Environmental turbulence Rapid and unpredictable change in both competitors' offerings and in customers' needs.

Functional business model The interaction of functions within a business.

Hypermedia Electronically based media that allow hyperlinks, or the ability to click on a link to new information. A prime example is the World Wide Web.

Lifetime value of customers The process of viewing the customer as a stream of future revenue rather than as a single transaction. This allows the identification of customers who provide value to the firm.

Margin The difference between the potential value delivered and the cost of creating that value.

Market of one The process of treating each customer as an individual. Often this is made possible through the use of data collected in databases and mined for information.

Online marketplaces (e-marketplace) A centralized online location for the trading of information, goods, services, or other commodities.

Primary activities Inbound logistics, operations, outbound logistics, marketing and sales, and service.

Support activities Procurement, technology development (including R&D), human resource management, and infrastructure (systems for planning, finance, quality control, information management, etc.).

Value chain analysis process The process of disaggregating the activities of buyers, suppliers, and the firm into a chain of interrelated value creating activities.

Value chain A way of envisioning the collection of activities that a business undertakes in order to design, produce, market, deliver, and support products or services.

Value web The horizontal and vertical relationships that exist in competitive systems, in which alliances can be virtual and arrangements temporary.

Virtual value chain A value chain framework where a firm captures information along its entire value chain and then uses that information to enhance value and improve customer relationships, or sells the information.

CROSS REFERENCES

See *Business-to-Business (B2B) Internet Business Models; Business-to-Consumer (B2C) Internet Business Models; Electronic Commerce and Electronic Business; E-marketplaces; Supply Chain Management.*

REFERENCES

Caldwell, B. (1999, February 8). Time and money pay off. *Information Week*, p. 16ER.

Caldwell, B., & Wilde, C. (1998, June 29). Emerging enterprises. *Information Week*, 53–60.

Carr, D. F. (2001, June 15). Forging 21st-century value chains. *Internet World*, 26–32.

Cartwright, S. D. & Oliver, R. W. (2000, January–February). Untangling the value web. *The Journal of Business Strategy*, 22–27.

Chabrow, E. (2000, March 6). Seeking the deeper path to e-success. *Information Week*, 49–76.

Chadderdon, L. (1998, September). How Dell sells on the Web. *Fast Company*, 78–88.

De Soto, R. (1998, December). Creating an active Internet presence: A new alternative. *Telecommunication Magazine*, 73–75.

Downes, L., & Mui, C. (1998). *Unleashing the killer app: Digital strategies for market dominance*. Boston, MA: Harvard Business School Press.

Fingar, P., & Aronica, R. (2001, June 15). Empower your customers—The driving forces of the real new economy. *Internet World*, 33–35.

Jones, K. (2003, February). The Dell way. *Business 2.0*, 60–66.

Kleindl, B. (2000). Competitive dynamics and opportunities for SMEs in the virtual marketplace. *Journal of Developmental Entrepreneurship, 5*(1), 73–85.

Kleindl, B. (2003). *Strategic electronic marketing: Managing e-business*. Cincinnati, OH: South-Western College Publishing.

Kuttner, R. (1998, May 11). The Net: A market too perfect for profits. *Business Week*, 20.

Magretta, J. (1998, March–April). The power of virtual integration: An interview with Dell Computer's Michael Dell. *Harvard Business Review*, 72–84.

Porter, M. E. (1985). The value chain and competitive advantage. In *Competitive advantage: Creating and sustaining superior performance*. New York: Free Press.

Porter, M. E. (2001, March). Strategy and the Internet. *Harvard Business Review*, 62–78.

Porter, M. E., & Millar, V. E. (1985, July–August). How information gives you competitive advantage. *Harvard Business Review*, 1–13.

Rayport, J. F., & Soviokla, J. (1995, November–December). Exploiting the virtual value chain. *Harvard Business Review*, 75–85.

Rosenau, M. D., Jr. (1990). *Faster new product development*. New York: AMACOM.

Serwer, A. (2002, January 21). Dell does domination. *Fortune*, 70–84.

Slater, D. (1998, December 15–1999, January 1). The corporate skeleton. *CIO*, 100–106.

Szygenda, R. (1999, February 8). Information's competitive edge. *Information Week*, 4ER–10ER.

Wilder, C. (1998, June 29). Internet levels the field. *Information Week*, 64–66.

Wilder, C. (1999, January 4). E-business work status. *Information Week*, 53–54.

FURTHER READING

Clayton, C. (2000). After the gold rush: Patterns of success and failure on the Internet. Retrieved January 2, 2002, from Innosite.com Web site: http://www.innosight.com/#After%20the%20Goldrush%20Free.PDF

Engler, N. (1999, January 18). Small but nimble. *Information Week*, 57–62.

Varian, H., Litan, R. E., Elder, A., & Shutter, J. (2001, December 10). *The Net impact study: Preliminary report*. Retrieved July 30, 2002, from Netimpactstudy.com Web site: http://www.netimpactstudy.com/

Video Compression

Immanuel Freedman, *Dr. Immanuel Freedman, Inc.*

INTRODUCTION

Definition of Video Compression

Video compression is a process of reducing the amount of digital data used to represent a sequence of images normally varying in time and intended to portray motion subject to the requirements that the quality of the reconstructed video is sufficient for a certain application and the complexity of the computation involved is appropriate for that application.

Some researchers, for example, Shi and Sun (2000), use the term *video* to refer exclusively to image frames and sequences associated with the visible band of the electromagnetic spectrum while others, for example, Tekalp (1996), refer exclusively to sequences.

Example of Video Compression

Figure 1a shows images of frames numbered 1231–1233 from "The Emotion of Space" movie (NASA, 2001) indicating a scene change at frame 1232 and motion sequence from frames 1232 to 1233. Figure 1b displays a graph of the data rate and sample size by frame number centered on frame 1231 for the aforementioned movie. With an image size of 320 × 240 pixels, pixel depth of 24 bits (8 for each of the red, green, and blue channels), and an average frame rate of 15 frames per second (fps) to create an illusion of motion by the persistence of vision, the 3219 frames of this movie of about 215 s duration would require a data rate of about 27.65 Mbps to be transmitted in real time without compression, not considering the audio tracks, far above even the capacity of a 1.5 Mbps T1 communications line with a download time for the 742 MB data exceeding 8 min under optimal conditions.

Objective of Video Compression

Video compression provides a technology solution for applications in which the required data rate for communication, manipulation, or storage of digital video sequences exceeds the capacity of communications channels or storage devices. Table 1 lists the data rate and corresponding compression ratio required for several applications in common use.

Compression Paradigms

Video compression methods are often described as "lossless" or "lossy." Lossless compression methods are reversible unlike lossy methods, which comprise irreversible changes to data. The term lossy arose from an analogy with a concept familiar to electrical engineers, the dissipation of electrical energy as heat in the transmission of power. The approximations to data made by lossy compression methods frequently yield far greater compression than the exact lossless methods for a small reduction in information content. In hybrid schemes, residuals from the approximations made by a high compression (perhaps 100:1) lossy method can, in turn, be coded by a moderate compression (perhaps 2:1) lossless method to yield an exactly reversible combination of compression methods. This allows the user to browse a highly compressed approximation to the data and make precise calculations with an exact reconstruction when required. The terms "approximate" and "exact" may be more acceptable than the terms lossy and lossless to scientists and administrators, who often take the stance that "loss is intolerable" but "an approximation is good enough."

(a)

(b)

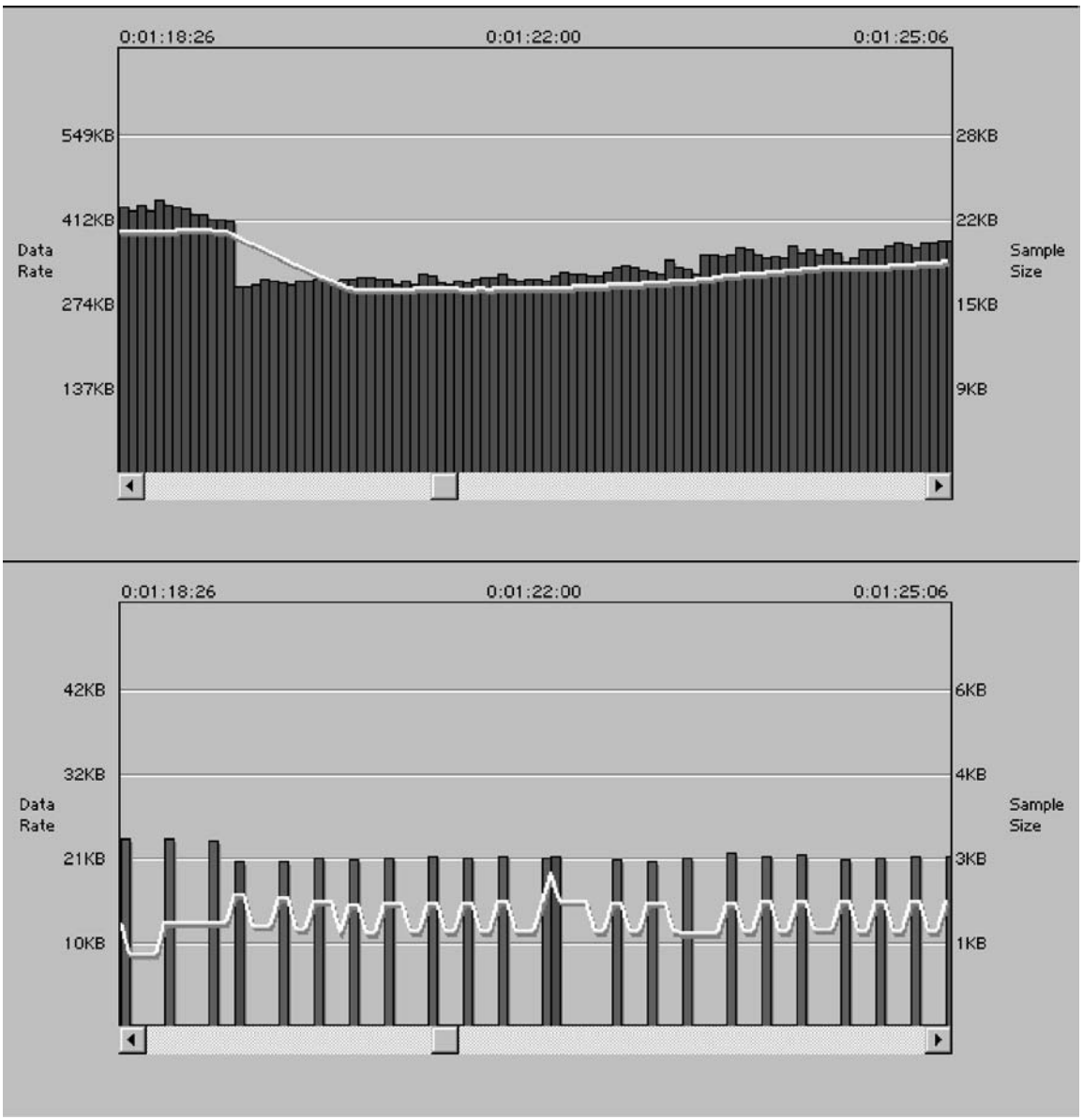


Figure 1: (a) Images of frames numbered 1231–1233 from the NASA budget FY 2001 movie "The Emotion of Space" showing scene change at frame 1232 and motion sequence from frames 1232 to 1233. (b) Graph showing data rate and sample size by frame centered at time code 0:01:22:00 (frame 1231) for H.263 compression of the above movie at 15 fps and average data rate 293 and 14 kbps, respectively.

Table 1 Data Rate and Corresponding Compression Ratio Required for Several Well-Known Applications

Application	Network	Data Rate	Video Parameters	Compression Ratio
Mobile customer service video	iMode	About 9.6 kbps	About 320 × 240 × 16 at 15 fps	About 1920:1
Video-conferencing, Shopping	POTS	About 56 kbps	About 320 × 240 × 16 at 15 fps	About 329:1
	ISDN, FOMA (initial)	About 64 Kbps	About 320 × 240 × 16 at 15 fps	About 288:1
	GPRS	About 128 Kbps	About 320 × 240 × 16 at 15 fps	About 144:1
	FOMA (planned)	About 384 kbps	About 320 × 240 × 16 at 15 fps	About 48:1
	UMTS	Up to 2 Mbps	Up to 1024 × 768 × 24 at 29.97 fps	About 566:1
Digital Movie	Broadband	About 500 kbps–1.5 Mbps	Up to 1024 × 768 × 24 at 24 fps	About 905:1
Digital Television	Cable HFC	About 1.5–4 Mbps	Up to 1080 × 1920 × 24 at 60 Hz (29.97 fps)	About 746:1
Digital Versatile Disk	DVD distribution	About 6–10 Mbps	Up to 720 × 480 × 24 at 29.97 fps	About 42:1
Digital Cinema	Satellite	About 60–80 Mbps over 45 Mbps line	Up to 2560 × 1080 × 30 at up to 29.97 fps	About 55:1

Operations and Infrastructure Supported by Digital Video

Video compression facilitates system functions comprising storing, retrieving, transmitting, receiving, nonlinear editing, indexing, manipulating, restoring, and denoising digital video sequences.

Video compression technology provides a significant component of infrastructure for familiar systems including digital, interactive, or time-shifted television, video-on-demand, home and in-flight entertainment or shopping, interactive games and multimedia, streaming media, video editing suites, camcorders, security cameras, and medical scanners.

DIGITAL VIDEO SIGNAL REPRESENTATION

Anatomy of a Digital Video Sequence

A motion picture scene is a subsequence of images (frames) usually resulting from continuous camera action. The images that define a digital video sequence comprise an ordered set of pixels. A pixel may be defined as the smallest distinguishable and resolvable area in an image. The number associated with an image pixel may refer to any information of interest including sampled measurements of light, temperature, image feature location, or visual object trajectory. The pixels of multidimensional images are labeled by at least two quantities referred to as dimensions. A hyperspectral image records a collection of subimages each of which is gathered from a distinct region of the electromagnetic spectrum. An inhomogeneous image records a collection of diverse information. An ensemble is a collection of images described in statistical terms.

Sources of Digital Video

A digital video sequence may be captured or created directly in digital form or converted from an analog video source. In view of the widespread use of existing analog display devices, digital video sequences are frequently converted to analog form and current digital television standards are expressed in terms consistent with existing analog television systems. International Telecommunication Union (ITU) Recommendation ITU-R BT.470–6 (11/98) records characteristics of conventional analog television systems including the National Television Standards Committee (NTSC), phase alternating line (PAL), and Systemè Electronique Couleur Avec Memoire (SECAM) adopted by a number of countries worldwide. Although discussion of analog video input types is outside the scope of this article, they may be classified as composite, component analog, or S-Video. In contrast, a serial digital interface provides connections for ITU-R BT. 601–5, composite digital video, and four channels of digital audio with a transfer rate of about 270 Mbps.

Interlaced Digital Video

According to ITU-R BT.470–6, all the television systems listed in Annex 1 of that recommendation employ an aspect ratio of the picture display (width/height) of 4/3, a scanning sequence from left to right and from top to bottom, and an interlace ratio of 2/1 resulting in a picture (frame) frequency of half the field frequency. Odd and even scan lines are traversed on successive scans.

Noninterlaced (Progressive) Digital Video

In contrast with the Recommendation ITU-R BT.470–6, noninterlaced (progressive) video sequences result from scanning all the lines of a picture frame once per field resulting in a picture (frame) frequency equal to the field frequency. In principle, video frames may be scanned in any order such as a depth-first tree search widely used for tiled images.

Color Models

The Commission Internationale de l'Eclairage (CIE) defines standards for color illuminants and observers

including the chromaticity coordinates diagram defined in a joint ISO/CIE standard ISO/CIE 10527:1991, which supersedes CIE S002(1986). The range (gamut) of colors visible to the human eye is described by the LAB gamut in the CIE chromaticity diagram, larger than the EKTA film gamut which is, in turn, larger than the RGB digital media gamut.

Poynton (1999a,b) described aspects of color specification and coding relevant to video signal processing in the form of Frequently Asked Questions available online. The CIE home page (Commission Internationale de l'Eclairage, 2000), the HyperPhysics Web site (Nave, 2001), and the Color Metric Converter Web pages (Dawes, 1999) serve to encourage further online exploration of human vision and color concepts.

The analog NTSC television standard SMPTE 170M-1999 specifies the system reference white and the primary color channels (green, blue, red) in terms of chromaticity coordinates The system reference white has been chosen to match the chromaticity coordinates of the D_{65} illuminant defined by joint ISO/CIE standards ISO/CIE ISO 10526:1999 and CIE S 005 E-1998 intended to represent average daylight with a correlated color temperature of approximately 6500K.

According to SMPTE 170M-1999 section 5, a cathode ray tube display has an inherently nonlinear electro-optical transfer characteristic and to achieve an overall system characteristic that is linear, it is necessary to specify compensating nonlinearity elsewhere in the system. In NTSC, PAL, and SECAM systems, the signal is pre-corrected for non-linearity at the signal source assuming the display is intended to be viewed by human observers in a dimly-lit environment. Although Table 1 of ITU-R BT.470–6 (conventional television systems) lists the assumed gamma of the display device for which a pre-correction of a monochrome signal is made to be typically 2.2 (in some countries 2.8), the operating values of the respective transfer characteristics may vary from the precise values given in sections 5.1 and 5.2 of the standard to meet operational requirements in practical systems.

ITU-R BT.601–5 defines a method in section 3.5 for constructing digital video luminance and color difference signals (Y, C_R, C_B) comprising the steps of constructing analog luminance E_Y and color difference signals ($E'_R - E_Y$, $E'_B - E'_Y$) as weighted sums and differences of the corresponding gamma-corrected analog television color signals (E'_R, E'_B, E'_G), renormalizing the analog television luminance and color difference signals to the ranges

$$E'_Y \in [1.0, 0.0], \quad (E'_R - E'_Y) \in [+0.5, -0.5],$$
$$(E'_B - E'_Y) \in [+0.5, -0.5],$$

and quantizing these renormalized analog luminance and color differences signals (chrominance) to a uniformly-quantized 8-bit binary encoding equivalent to a decimal range [0, 255] to provide digital luminance and color difference signals (Y, C_R, C_B). The digital luminance occupies only 220 levels to provide working margins with black at level 16 and the color difference signals occupy 255 levels with zero value at level 128. It is usual to limit the gamut of digitally coded (Y, C_R, C_B) signal value to that supported by the corresponding ranges of (R, G, B) signals.

Sampling, Quantizing, and Coding

We conventionally think of images as sampled and quantized at a sampling rate large enough to preserve the useful information. When we sample light measurements, we usually record both the brightness (luminance) and color (chrominance) information associated with an image pixel. In a motion picture application we often record the frame number (or time code) to indicate the position or presentation time of the frame relative to the start of the scene or video image sequence. The frame aspect ratio of a rectangular image frame is the ratio of its width to height.

International Telecommunications Union Recommendation ITU-R BT.601–5 (formerly known as CCIR 601) specifies methods for digitally coding standard 4:3 and wide-screen 16:9 television video signals for both 525-line and 625-line television systems. In particular, the *pixel aspect ratio* (defined as the ratio of pixel height to width) may be more or less than 1/1 resulting in rectangular (*non-square*) rather than square pixels. Aho (2002) showed how to deduce the ratio 11/10 for rendering data from 525-line systems and 54/59 for 625-line systems from analog standard video sampling rates.

ITU BT.601–5 recommends co-siting the samples representing digital luminance and color difference signals (Y, C_R, C_B) to facilitate the processing of digital component video signals and further recommends subsampling to reduce the data rate. Horizontal subsampling of the color difference signals by 2:1 yields a 2/3 saving in data rate over (R, G, B) with almost imperceptible change in visual quality when implemented with proper decimation and interpolation filters. The 4:2:2 nomenclature indicates 4 luminance samples for every 2 color difference samples in a scan line. Horizontal and vertical subsampling by 2:1 is denoted by the 4:2:0 and yields about 1/2 the data rate for (R, G, B). In turn, 4:1:1 denotes horizontal subsampling by a factor of 4 without vertical subsampling. If a fourth parameter is mentioned—as in 4:4:4:4—it refers to the alpha channel required for keying applications. Note that digital luminance levels 0 and 255 are reserved for synchronization in 4:2:2 systems while levels 1 to 254 are available for video. Figure 2a illustrates the digital sampling of luminance and chrominance. Each cell indicates a luminance sample while chrominance samples are indicated by cells designated by "C" and alpha channels by cells designated by "A." The 4:4:4:4 sampling scheme samples chrominance and alpha channels with every luminance sample. The 4:2:2, 4:1:1, and 4:4:4:4 schemes sample chrominance with luminance samples while the 4:2:0 scheme samples chrominance between luminance samples.

ITU-R BT.601–5 digital video is sampled at 13.5 million pixels per second (for both 525- and 625-line systems). A 525-line analog NTSC (ANSI/SMPTE 170M-1994) video signal is sampled at exactly $12 + 27/99$ million pixels per second when sampling with square pixels. A 625-line analog PAL (ITU-R BT.470–3) video signal is sampled at

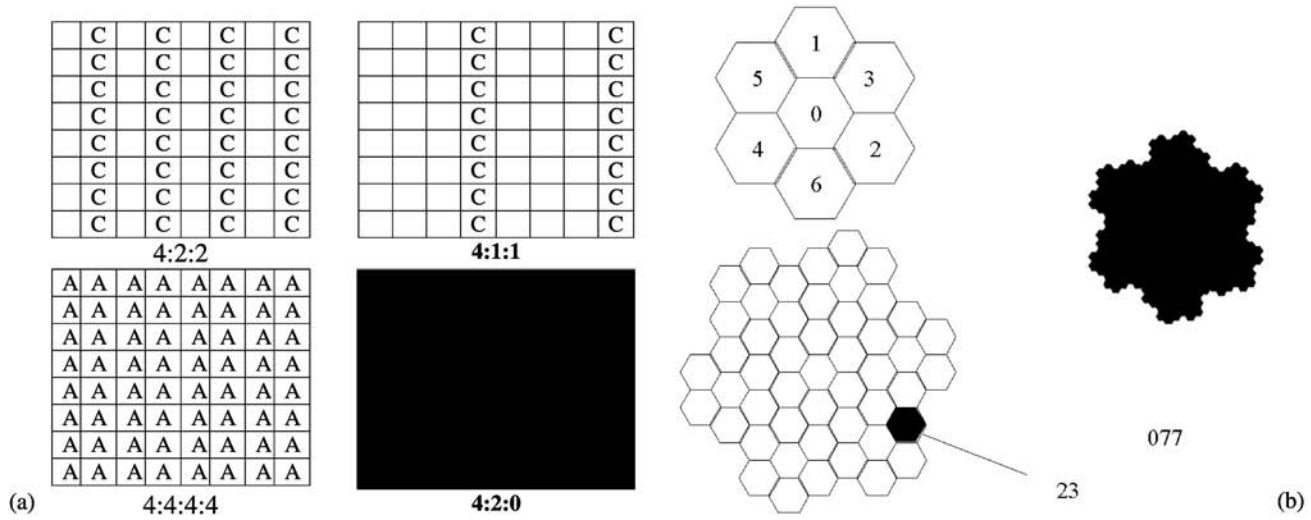


Figure 2: (a) Digital sampling of luminance and chrominance. Each cell indicates a luminance sample while chrominance samples are indicated by cells designated by "C" and alpha channels by cells designated by "A." The 4:4:4:4 sampling scheme samples chrominance and alpha channels with every luminance sample. The 4:2:2, 4:1:1, and 4:4:4:4 schemes sample chrominance with luminance samples while the 4:2:0 scheme samples chrominance between luminance samples. (b) Generalized Balanced Ternary Sampling and Addressing Scheme. Seven hexagons, labeled 0 to 6 form an aggregate, which can be used to tile the plane. The address "23" refers to the hexagon in position 3 of the aggregate in position 2. If we define the address 7 to refer to an aggregate of hexagons, the address "077" indicates a pattern formed from the aggregation of aggregates of hexagons.

exactly 14.75 million pixels per second. Although the ITU-R BT.601–5 Recommendation states that the duration of the digital active line for either 525-line or 625-line video signals is 720 samples, the active line is not clearly defined. Tables 1 to 6 of SMPTE RP187–1995 recommend values to quantify the center, aspect ratio and blanking of video images relative to a reference image lattice of width 1920 and height 1080 pixels, with derived pixel aspect ratios that Chris Pirazzi points out in his "Lurker's Guide to Video" to be somewhat impractical and inconsistent with current industry practice. A "full raster" image, coded at 10 bits per sample, contains timing, synchronization, closed-captioning, and other ancillary information.

Time Code

SMPTE 12M-1999 defines a time and control code for television, audio, and film systems operating at 29.97, 25, and 24 frames per second. A video frame often contains a time code similar to hours, minutes, and seconds as an index for editing purposes. According to section 4 of SMPTE 12M-1999, "the exact field rate for NTSC television systems is 60/1.001 fields per second (59.94 Hz) with color frame rate 29.97 Hz so that straightforward counting at 30 fps will yield an error of approximately +108 frames in an hour." "Drop frame time code" is an SMPTE time code option that allows the time code to agree with clock time. In drop-frame time code mode, frames 0 and 1 are dropped on every minute with this correction negated at the beginning of every tenth minute so that the time code agrees with clock time as if the frame rate were truly 30 fps. Note that 25-frame systems (PAL) for which the frame rate is exactly 50 Hz and 24-frame systems (film media) do not require a drop frame option.

Telecine, Pull-Down, and Inverse Telecine

Teranex (Teranex, 2002) pointed out that the conversion of 24 fps film material to 29.97 fps digital video (telecine) takes place through a process commonly known as 2:3 (or 3:2) pull-down in which four consecutive film frames are padded out to five digital frames of 10 digital fields. An encoder will often try to detect that it is receiving film-originated material and remove the redundant fields in an inverse telecine process to provide additional compression. However, the motion and scene changes in the original material together with the editing post-processing of digital video can often lead to corrupt 2:3 sequences causing the encoder to spend additional resources on inverse telecine. This can be corrected by pre-processing the material to create a correct 2:3 cadence.

Tiling Techniques

In principle, a pixel can be any suitable shape, not just square or rectangular. Grunbaum and Sheperd (1980) classified the 81 isohedral tilings of the plane. Herring (1994) of the U.S. Army Construction engineering Research Laboratory pointed out that hexagonal tiles have uniform adjacency with six nearest neighbors and may be arranged hierarchically to establish a scalable domain-specific software architecture model for mapping virtual environments onto network computing resources. Herring (1994) showed how a hexagonal tessellation organizes space into a hierarchy.

Addressing Techniques

Gibson and Lucas (1982) pointed out that the tiles of the aforementioned hexagonal scheme may be organized into rotational hierarchies of seven hexagons and addressed via a hierarchical place value system known as Generalized Balanced Ternary, occasionally used in image and

vision processing. Figure 2b shows the Generalized Balanced Ternary sampling and addressing scheme. Seven hexagons, labeled from 0 to 6, form an aggregate that can be used to tile the plane. The address "23" refers to the hexagon in position 3 of the aggregate in position 2. If we define the address 7 to refer to an aggregate of hexagons, the address "077" indicates a pattern formed from the aggregation of aggregates of hexagons.

See Snyder, Qi, and Sander (1999) and the exploration of tesseral addressing schemes advocated by Bell, Diaz, Holroyd, and Jackson (1983), and Jackson, Bell, Stevens, Freedman, and Dickman (1986).

DIGITAL VIDEO SIGNAL COMPRESSION

Overview and Tradeoff Dimensions

The video compression process typically trades communications or storage bandwidth for processor cycles and picture quality determined by the intended application of the digital video sequence. In addition to considerations of image compression applied to each frame independently of the others (intraframe coding), scenes showing moderate motion may be effectively compressed by considering what has changed between frames (interframe coding and conditional replenishment). A motion picture will often have video sequences synchronized with audio tracks and ancillary information such as system parameters. A video sequence may be coded at multiple bit rates or in multiple versions.

Digital Video Quality Assessment

Measuring the quality of a video sequence reconstructed from compressed imagery is an application-dependent task. While there may be well-defined criteria for acceptability of the compressed video sequence intended for machine use in a scientific or medical discipline, considerable research activity is still needed to close the gap between objective and subjective assessments of quality for video sequences intended for human viewing.

Human Visual Response

Assessment of video quality by human observers, invited to view scenes encoded with varying parameters takes the human visual response (HVR) explicitly into account. Jain (1989) and the HyperPhysics Web site (Nave, 2001) introduced the following key elements of the HVR.

Eye movements focus scenes onto the retina, which contains rod and cone photoreceptors. The cones are clustered about the center of the retina (*fovea centralis*) and are sensitive to red, blue, and green color stimulus (*photopic vision*). Although the blue-sensitive cones number much less than the red- or green-sensitive cones, their sensitivity is far greater. The rods are highly sensitive to peripheral motion at low light levels and relatively insensitive to color (*scotopic vision*).

Humans perceive luminance in terms of relative contrast (Weber's Law).

The Mach Band effect demonstrates that perceived brightness is not a monotonic function of luminance (*lateral inhibition*).

MacAdam ellipses demonstrate regions of *just noticeable difference* in color in the CIE chromaticity diagram.

Bloch's Law states that "light flashes of different durations but equal energy are indistinguishable below a critical duration." This critical duration depends on how well the eye is adapted to the dark and is nominally about 30 ms.

When a light flashes at a rate above the critical fusion frequency, the flashes are indistinguishable from a steady light of average intensity.

The eye is more sensitive to flickering of high spatial frequencies than low spatial frequencies.

Human attention is attracted by faces in video scenes but may be distracted by peripheral object motion.

Subjective Evaluation

Subjective evaluation requires a group of human observers—preferably not expert in image quality assessment—to view and rate video quality in terms of a scale of impairments, which range from "not noticeable" to "extremely objectionable." ITU-R BT.500–10, Methodology for the Subjective Assessment of the Quality of Television Pictures, recommends a specific system prescribing viewing conditions, range of luminance presented to the viewer panel, number and experience of viewers, monitor contrast, selection of test materials, and process for evaluation of test results. The Double Stimulus Impairment Scale and the Double Stimulus Continuous Quality Scale are particularly noteworthy. Subjective tests are costly and not highly reproducible. ANSI T1.801.01–1995 provides a set of test scenes in digital format while ANSI T1.801.01–1996 provides a dictionary of commonly used video quality impairment terms.

Objective Evaluation

Objective evaluation techniques range from simple test metrics that do not take the HVR into account (such as the peak signal-to-noise ratio often quoted by researchers) to the vision system model metric developed by the Sarnoff Corporation (Sarnoff Corporation, 2001), which relies on comparison of maps of just noticeable differences between original and compressed video sequences, one frame at a time. Annex A of ANSI T1.803.03–1996 lists a set of objective test criteria that may be used to measure video quality in one-way video systems, applying objective tests closely related to known features of the HVR. Webster et al. (1993) presented a scheme for combining subjective and objective assessments on test scenes based on objectively generated impairments.

Rate Distortion Relationships

The Shannon (1948) rate distortion bound refers to the minimum average bit rate required to encode a data source for a given average distortion level. If the data can be perfectly reconstructed, the bit rate at zero distortion is equal to the source entropy, a measure of the information contained in the source. In practice, many

researchers compare the effectiveness of video compression algorithms by plotting empirically determined quality measures as a function of observed bit rate for a given set of test sequences. Almost invariably, the observed video quality improves with increased bit rate.

Principles of Digital Video Compression

Digital video compression operates by eliminating redundant spatial, temporal, hyperspectral, statistical, or pyschovisual information. A brief inspection of Figure 1a shows considerable overlap between frames 1232 and 1233 but a discontinuous change of scene (likely as a result of editing) between frames 1231 and 1232. All the frames shown exhibit strong local spatial correlation, the motion of the astronauts exhibits temporal correlation, the viewer's attention is focused on the motion of astronauts and the expressions on their faces, and the space suit color contrasts well with the background. Figure 1b illustrates the well-known video artifact of "dropped frames" when this movie is compressed at an average data rate less than the original 293 kbps.

Human Destination

Compression of video sequences intended for human viewing may take place by a combination of lossy and lossless compression steps. Local spatial correlation may be removed by intraframe coding techniques such as block matching or transform coding followed by quantization of the transform coefficients (see Data Compression).

Temporal correlation may be removed by interframe coding techniques predicting the motion vectors observed in differences between successive frames. Spectral correlation may be removed by applying temporal decorrelation or modeling techniques to the spectral dimensions of the hyperspectral imagery; coding redundancy may be removed through careful design of compression codes, while psychovisual redundancy may be addressed by techniques that drop frames and increase the bit rate in regions containing human faces, sharply contrasting regions and trajectories of distracting objects that move rapidly through the peripheral field of view.

Machine Destination

In scientific and medical applications, imagery may be correlated in any of the preceding ways. There are likely to be additional constraints on the compression which may include fidelity criteria (such as 90% of the encircled energy to remain within the pixel for remotely sensed land use data, or the peak difference in value reconstructed from the compressed imagery to be within 10% of the original value for 90% of the reconstructed pixels) together with processing constraints such as the bit rate must remain approximately constant within the communication channel capacity regardless of video quality or content and consume no more than 5% of the available processing resource.

As an example, Freedman, Boggess, and Seiler (1993) and Freedman and Farrelle (1996) set forth criteria for the experimental compression system developed to optimize real-time calculations on large diverse data sets from NASA's Cosmic Background Explorer mission in a clustered computing environment in which the application software exceeded 1M lines of code:

Provide compression transparently without changing the application software.

Compress instrument pipeline and science analysis data products to better than 16–50%.

Process compressed data at a worst case throughput not less than 90% of uncompressed data processing.

Preserve required accuracy of instrument housekeeping and scientific data according to specified validation criteria.

Exceed bitwise reliability of 10^{-13} on average (flawless compression of 380 GB). Several times this factor is desirable.

Support full random access to data records.

Provide a capability to select specific classes of data for compression.

Provide a capability to select a compression scheme (*representation*) for each field of a data record.

Preserve overlaps in separately processed data segments.

Store search keys (e.g., time code and pixel address) in plain codes.

Optimize choice of representation, combining a priori scientific knowledge with adaptive knowledge of data.

Provide mechanisms for easy user configuration and database management of compressed data.

Pre- and Postprocessing

Video source material is often pre-processed before encoding and post-processed before distribution. Common preprocessing steps include digital nonlinear editing (Ohanian, 1998), transcoding from another compressed representation, and the conversion of film-originated material from 24 to 29.97 fps via the telecine process just discussed. Common postprocessing steps include digital nonlinear editing, the inverse telecine process, and the manual deletion or recoding of specific frames.

Motion Estimation and Compensation

Strong temporal correlation between successive video frames suggest that the bit rate of a video sequence may be reduced by interframe coding methods such as conditional replenishment, motion-compensated coding, and three-dimensional transform coding.

The conditional replenishment approach seeks to encode only the difference in pixel values which change between successive frames and requires considerably less processing power and simpler algorithms than either motion-compensated coding or three-dimensional transform coding. Motion-compensated methods usually result in data rates below that resulting from conditional replenishment and comprise determining motion vectors from frames preceding the current frame by methods such as block matching, region matching, or analysis of optical flow, predicting the motion vectors for successive frames, estimating the prediction error of these motion vectors, and encoding the predicted motion vectors

together with the estimate of their prediction error (Shi & Sun, 2000). The three-dimensional transform technique has become feasible due to recent advances in electronics manufacture and represents a growing area for research.

Rate Control of Compressed Digital Video

The information content of digital video sources may vary in complexity from scene to scene. Constant bit rate encoding forces the reconstructed video quality to vary in order to maintain a (near) constant bit rate for the channel coder. Variable bit rate (VBR) has the goal of maintaining (near) constant quality by varying the bit rate allocated to scenes of varying complexity. Popular methods of rate control include varying the quantization of transform coding in a feedback loop and embedded zero tree encoding (Shapiro, 1993).

Psychovisual Modeling

Human attention is often attracted to faces and facial expressions and distracted by the motion of peripheral objects across the field of view (see Colmenarez, Frey, & Huang, 1999). An encoder may choose to encode these regions of interest at a higher bit rate than others if the regions can be bounded (typically, by focus boxes).

Statistical Multiplexing

Often, a broadcaster wishes to transmit several program streams on a channel. Statistical multiplexing is a process by which the instantaneous encoding rates of multiple program streams are adjusted by VBR techniques to maintain either (near) constant quality or optimize the aggregate bit rate in the channel.

Network Transmission Issues

In networks with quality of service guarantees such as the Internet, the varying availability of bandwidth leads to network congestion and packet jitter. Approaches for addressing these issues include buffering the video stream and employing techniques that yield multiple bit rates for same stream. Hunt (2001) filed a patent application that indicates a congestion management scheme—intended primarily for mobile videophone users—in which graphics and text may be displayed when it is necessary to buffer the video stream so that the user perceives the video streaming process to be in real time.

DIGITAL VIDEO COMPRESSION STANDARDS

MPEG-1, -2, -4 Visual Codecs

Introduction

Koenen (2001) pointed out that the Moving Picture Expert Group (MPEG)—whose formal name is ISO/IEC JTC1/SC29/WG11—has developed the widely used MPEG-1, -2 and -4 codecs to facilitate and standardize infrastructure for the interoperability of multimedia. MPEG-1 is still widely used on the Web, targeted at bit rates 64 kbs –1.5 Mbps, MPEG-2 is the basis for current digital television and Digital Versatile Disk (DVD) standards, targeted at 1.5 Mbps –7 Mbps or higher, and MPEG-4 is an extensible toolbox of algorithms that addresses the coding of audiovisual objects throughout an extended bit range from about 5 bps to more than 1 Gbps. The standards address video compression, together with audio and systems elements. The MPEG Web site (http://mpeg.telecomitalialab.com) provides much detailed information about the MPEG standards including an overview.

Intra-, Predictive, and Bidirectional Frames

Table 2 shows the relationship among intra (I), predictive (P), and bidirectional (B) frames referred to in the MPEG standards and introduced in the MPEG-1 standard (ISO/IEC 11172–2:1993). I frames are coded using information only from that frame. P frames are coded using forward prediction from previous I or P frames. B frames are encoded from previous and succeeding frames using I or P frames. This, in turn, implies that the encoding order may be markedly different from the display order. Since B frames require both past and future frames to be decoded before they can be decoded and displayed, the presence of B frames adds considerably to the encoder delay. The MPEG-2 simple profile at main level requires only I and P frames, thus reducing delay.

Sequence, Group of Pictures, Slice, Macroblock, Block, and Pixel

A group of pictures (GOP) contains I, and potentially P, or B frames. During encoding the user selects the duration of the GOP as the number frames between successive I frames (N) together with the distance between

Table 2 Display and Transmission of a Group of 18 NTSC MPEG-2 Intra (I)

Display	I	B	P	B	P	B	P	B	P	B	P	B	P	B	P	B	P	B
Frame	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Transmit	B	I	B	P	B	P	B	P	B	P	B	P	B	P	B	P	B	P
Frame	2	1	4	3	6	5	8	7	10	9	12	11	14	13	16	15	18	17

Note: Predicted (P) and bidirectional (B) pictures. The P pictures are predicted from the previous I picture and the B pictures are interpolated from both the previous and next I or P pictures. The decoder must buffer P frames and compute B frames. The maximum length of a group of pictures (GOP) is 18 pictures on the NTSC system and 15 pictures on the PAL system.

nearest I or P frames (M). Prediction errors are propagated throughout the GOP and accumulate until the next I-frame is reached. If N is large, the accumulated error may become unacceptable. If M is 1, the encoder will use only I and P frames.

A frame may be divided into slices, which contains rows of macroblocks, from a single macroblock row to a full frame. A macroblock contains the luminance information Y for a 2×2 block area together with the subsampled chrominance information C_B, C_R, which, for 4:2:0 subsampled video sequences, may cover as small as one block. Each block, in turn is defined to contain 8×8 pixels over which discrete cosine transform coding is performed (see Data Compression). MPEG-2 Test Model 5 rate control depends on adjusting the quantization of the transform coefficients resulting from all the blocks of a macroblock slice in a feedback loop.

MPEG-4 System Decoder

In the same way that MPEG-1 and -2 standardize only the bit-stream syntax and decoder algorithms, MPEG-4 standardizes a system decoder model. The Koenen (2002) overview of the MPEG-4 standard showed the concept of encoding scenes containing visual objects and sprites. The discrete cosine transform, quantization, inverse discrete cosine transform, inverse quantization (see Data Compression), shape coding, motion estimation, motion prediction, and motion texture coding all combine to optimize the visual layer of the video compression system for specific content.

Error Resilience

Video compression typically uses interframe compression techniques to optimize the bit rate. Furthermore, the channel coder receives variable length codes as input. Any transmission errors, to which channels without quality of service guarantees (such as the wireless Internet) are especially prone, may cause loss of synchronization and the inability to reconstruct certain frames. Error resilience techniques such as those defined in the MPEG-4 and H.263 standards serve to reduce this risk. Resynchronization markers are inserted periodically at the start of a video packet, based on a predetermined threshold number of encoded bits and reversible variable length codes allow some of the data between resynchronization markers to be recovered if the data have been corrupted by errors.

Error Concealment

Koenen (2002) pointed out that a data partitioning approach can improve on the simple expedient of copying blocks from a previous frame when errors have occurred. A second synchronization marker is inserted into the bit stream between the motion and the texture information. When errors occur, the texture information is discarded and the motion information used to compensate the previous decoded video packet. In real-time situations, where there is a backchannel from the decoder to the encoder, dynamic resolution conversion may be used to stabilize the transmission buffering delay.

Content-Based Compression

The MPEG-4 standard supports content-based coding, random-access to content objects, and extended manipulation of content objects.

Shape and Texture Coding

The shape-adaptive discrete cosine transform based on predefined orthonormal sets of one-dimensional discrete cosine transform functions (see Kaup & Panis, 1997) can be used to encode visual objects of arbitrary shape (not just rectangles) together with texture.

Sprite Coding

MPEG-4 supports syntax for the efficient coding of static and dynamically generated sprites, which are still images representing backgrounds visible throughout a scene of a video sequence.

Object Coding

MPEG-4 supports the coding of audiovisual objects. Figure 3 depicts an audiovisual scene containing scrolling text, audio, background sprite, arbitrarily shaped video, and graphics.

General Scalability

The MPEG-4 Standard supports the scalability of visual objects through the following profiles described in Table 3.

MPEG-7 Visual and MPEG-21 Standards

The MPEG-7 multimedia content description interface (Martinez, 2001) supports query-by-content via descriptors expressed in the extensible markup language. In the visual case, the standard intends to convey basic and sophisticated information about the color, texture, shape, motion, localization of visual objects, and the recognition of faces.

The stated vision for the MPEG-21 multimedia framework (Bormans & Hill, 2000) is "to enable transparent and augmented use of multimedia resources across a wide range of networks and devices." The standard is intended to support interoperable content representation and intellectual property rights management in a "scalable and error resilient way. The content representation of the media resources shall be synchronisable and multiplexed and allow interaction."

MPEG-4 includes hooks for an open intellectual property management and protection scheme. A more interoperable solution is planned for development in the MPEG-21 standard. Bormans and Hill (2000) further described specific interactions and showed how the MPEG-7 standard supports transactions that produce and consume digital data items.

ITU-T Visual Codecs

The ITU visual codecs H.261 and H.263, Video Codec for Audio Visual Services at px64 kbit/s and Video Coding for Low Bit-Rate Communications, are primarily used for video conferencing applications, in which data rate and end-to-end delay are important for lip synchronization, a situation not encountered in broadcast applications. However, MPEG-4 contains many concepts derived

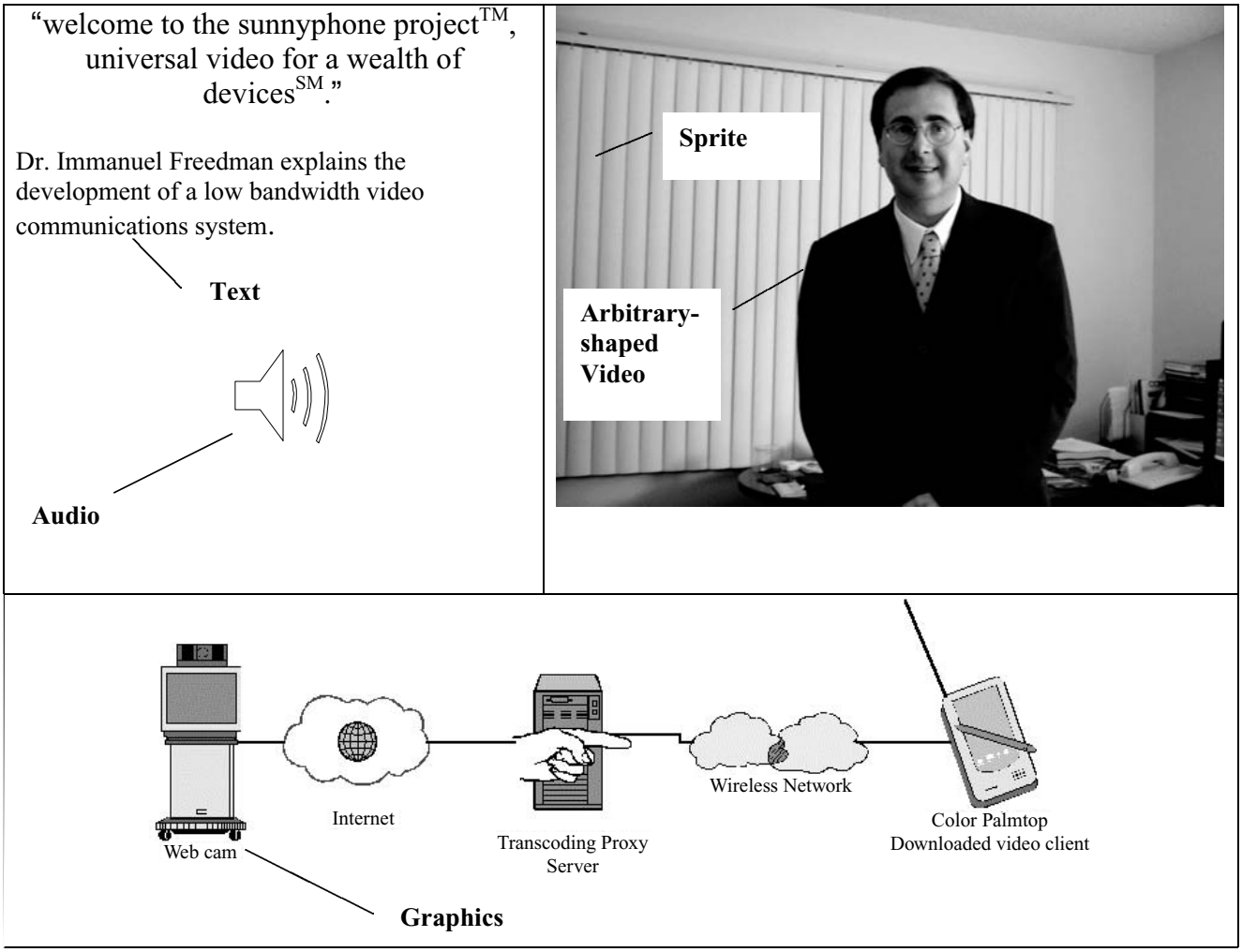


Figure 3: Audiovisual scene containing scrolling text, audio, background sprite, arbitrarily shaped video, and graphics.

from H.263. ITU-T H.261 operates on a common intermediate format (CIF) source specified by section 3.1. CIF has an aspect ratio of 4:3 with 352 × 288 pixels and 2:1 horizontal and vertical decimation of the color difference signals. Quarter CIF (QCIF) further decimates the spatial extent of the CIF picture. ITU-T H.261 supports a capability to reduce the transmitted frame rate by skipping zero to three frames according to some external decision.

To reduce encoding delays, ITU-T H.261 supports intracoding macroblocks rather than pictures, requires

Table 3 MPEG-4 Standard Profiles

Profile	Objects	Use
Simple Visual	Rectangular video objects	Mobile networks
Simple Scalable Visual	Temporal and spatial scalable objects	Internet
Core Visual	Arbitrarily shaped and temporally scalable	Internet multimedia
N-bit Visual	Core visual, but 4–12 bits pixel depth	Surveillance
Scalable Texture Visual	Spatial scalable coding of still (texture) objects	Games, still cameras
Core Scalable	Temporal and spatially scalable arbitrarily shaped objects	Internet, mobile, broadcast
Advanced Scalable Texture	Arbitrarily shaped texture and still images	Image browsing
Advanced Core	Arbitrarily shaped video objects and still image objects	Interactive multimedia streaming over Internet
Fine Granularity Scalability	Truncation of bit stream at any bit position to adapt delivery quality as required	Any

intracoding of at least one macroblock in every 132. The picture structure is based on the Group of Blocks, I and P frames. According to section 4.2.2:

> Each picture is divided into groups of blocks (GOBs). A group of blocks (GOB) comprises one twelfth of the CIF or one third of the QCIF picture areas (see Figure 6). A GOB relates to 176 pels by 48 lines of Y and the spatially corresponding 88 pels by 24 lines of each of C_B and C_R.

ITU-T H.261 supports motion compensation at only full pixel resolution. ITU-T H.263 improves upon ITU-T H.261 by using half-pixel precision for the motion compensation and provides support for additional negotiable coding options for improved compression performance. A new ITU-T standard (tentatively named H.263+) is being developed in cooperation with the MPEG committee.

Proprietary Codecs

Cinepak (http://www.cinepak.com) and Indeo from Intel (http://www.intel.com) are in common use in streaming media applications. The Aware MotionWavelets codec (http://www.aware.com) is a wavelet-based codec using only intra-frame compression. Qubit from Quvis (http://www.quvis.com) is a three-dimensional wavelet codec, which provides both spatial and temporal compression. Note, in passing, that the wavelet-based codecs soften the brightness edges in a manner that some viewers and video editors find objectionable. These issues

Table 4 Additional Well-Known Applications of Video Compression, Together with Requirements and Possible Solutions

Application	Requirements	Possible Solutions
Video Conferencing (PSTN)	352 × 288 pixels at 29.97 fps, 24-bit color Compress about 73 Mbps to 56.6 kbps in real time (about 1289:1)	H.261, H.263, MPEG-4, Aware wavelet I frames
Desktop Video Editing	1024 × 768 pixels at 29.97 fps, 24-bit color Store 120-min movie on 18-MB disk Compress about 468 to 18 MB (26:1)	Cinepak, Indeo, MPEG-4, Aware wavelet I frames
Digital Cinema	1920 × 1080 pixels at 24 fps, 30-bit color Transmit movie to theater over 45 Mbps line Compress 1.5 Gbps to 45 Mbps (33:1)	Qualcomm ABS DCT (http://www.qualcomm.com) (Morley, 2000), Qubit (http://www.qubis.com)
Wireless Video	One- and two-way scalable real-time video from 160 × 120 pixels monochrome to 640 × 480 with 24-bit color over network without QoS at 9.6 kbps (iMode) and up to 2 Mbps (Universal Mobile Telecommunication System). Display video on small screens with low resolution and with 33–200 MHz requiring low power consumption for extended battery life.	MPEG-4 Visual with Dynamic Resolution Conversion, MPEG-4 Facial Animation wavelet video (http://www.packetvideo.com, http://www.envivio.com, http://www.drfreedmaninc.com)
Home Shopping	Customer sees and hears personal sales assistant on television and then customer speaks via telephone to assistant in call center equipped with camera, encoder telephone, call routing, and graphics workstation.	MPEG-2 (Simple Profile), MPEG-4 (http://www.iseetv.net, http://www.avaya.com)
Science	When the Galileo spacecraft High Gain Antenna failed to deploy close to Jupiter, the mission required compression of science and engineering data streams from 13.4 kbps to 10 bps for transmission via the S-band antenna.	See Cheung and Tong (1993) noting particularly the low complexity encoding on a small onboard processor. The encoded imagery could be unpacked with higher computing power on Earth. This situation is also relevant to two-way wireless video.
Medicine	Remote medicine and telesurgery—Full motion virtual reality stereo video compression over fiber optic, Internet 2, or satellite links (subject to delays of up to a few seconds)	See Riva, Gamberini, and Davide (2001) and a simulator at http://synaptic.mvc.acc.uk/home.html.

require further research. Macromedia Flash Animation (http://www.macromedia.com) provides vector graphics animation suitable for cartoon applications. Cartoons may be arbitrarily spatially scalable without loss of quality.

Qualcomm's Adaptive Block Size Discrete Cosine Transform uses a combination of quadtree and discrete cosine transform methods (see Data Compression) to encode high quality imagery for digital cinema applications. Lastly (Fisher, 1995), pointed out a number of mechanisms for encoding video sequences via fractal compression methods.

How Standards Are Defined and Described

When an industry sector perceives the need for a new standard, it communicates this information to the national standards body, which in turn proposes it as a

Table 5 Hypothetical Broadband Video-on-Demand Business Model[1]

Digital Video Business Model						
Broadband VoD Revenue	**Year 1**	**Year 2**	**Year 3**	**Year 4**	**Year 5**	**Year 6**
Avg Subs	10,297,000	13,248,000	16,410,000	19,491,000	22,535,000	24,989,000
% who use VoD	**2.4%**	**2.4%**	**2.4%**	**2.4%**	**2.4%**	**2.4%**
Avg users per day	247,128	317,952	393,840	467,784	540,840	599,736
Conversion Rate (daily)	**6.7%**	**6.0%**	**5.4%**	**5.0%**	**4.4%**	**6.7%**
purchases/day	16,558	19,173	21,110	23,506	23,554	40,182
Average Purchase	**$4.00**	**$4.00**	**$4.00**	**$4.00**	**$4.00**	**$4.00**
Monthly Subscription per Sub	**$8**	**$8**	**$8**	**$8**	**$8**	**$8**
Gross purchase revenue per day	$66,230.30	$76,690.02	$ 84,439.30	$94,024.58	$94,214.33	$160,729.25
Gross subscription revenue per day	$65,900.80	$84,787.20	$105,024.00	$124,742.40	$144,224.00	$159,929.60
Gross revenue per day	$132,131.10	$161,477.22	$189,463.30	$218,766.98	$238,438.33	$320,658.85
days in year	**365**	**365**	**365**	**365**	**365**	**365**
Annual Gross Revenue	$48,227,853	$58,939,186	$69,154,103	$79,849,949	$87,029,990	$117,040,480
Content Partner Revenue Share	**52%**	**52%**	**52%**	**52%**	**52%**	**52%**
Content Partner Net Revenue	$25,078,484	$30,648,377	$35,960,134	$41,521,974	$45,255,595	$60,861,049
Service Provider Avg Revenue Share	**48%**	**48%**	**48%**	**48%**	**48%**	**48%**
Service Provider Net Revenue	**$12,037,672**	**$14,711,221**	**$17,260,864**	**$19,930,547**	**$21,722,685**	**$29,213,304**
Capital Costs						
Maximum Concurrent Usage	**10%**	**10%**	**10%**	**10%**	**10%**	**10%**
New Streams Required	24713	31795	39384	46778.4	54084	59973.60
Server Capacity (bps)	200,000,000	200,000,000	200,000,000	200,000,000	200,000,000	200,000,000
Bandwidth per stream (kbps)	580,000	580,000	580,000	580,000	580,000	580,000
Number of New Servers	72	92	114	136	157	174
Cost per Server	$10,000.00	$4,996.17	$2,496.17	$1,247.13	$623.09	$311.31
Cost per Stream	$50.00	$50.00	$50.00	$50.00	$50.00	$50.00
New Server Costs	$716,671.20	$460,677.48	$285,097.01	$169,182.57	$97,727.43	$54,143.36
New Stream Costs	$1,235,640.00	$1,589,760.00	$1,969,200.00	$2,338,920.00	$2,704,200.00	$2,998,680.00
Total Server Costs	$1,952,311.20	$2,050,437.48	$2,254,297.01	$2,508,102.57	$2,801,927.43	$3,052,823.36
Amortized Server Costs	**$555,214.24**	**$1,177,269.74**	**$1,890,689.14**	**$2,292,285.65**	**$2,683,311.14**	**$2,893,965.57**
Operating Costs						
Bandwidth Costs						
Purchases per year	6,043,515	6,997,965	7,705,086	8,579,743	8,597,057	14,666,544
Average 120-min movie (MB)	4176	4176	4176	4176	4176	4176
Total Usage (MB)	25,237,719,642	29,223,499,936	32,176,438,134	35,829,007,979	35,901,311,828	61,247,487,243
Bandwidth costs ($ per MB)	5.E-05	8.E-06	1.E-06	2.E-07	3.E-08	4.E-09
Total Bandwidth Costs	**$1,261,886**	**$222,924**	**$37,447**	**$6,362**	**$973**	**$253**
Cost per Rack Unit	$900	$900	$900	$900	$900	**$900**
Rack Units (2RU servers)	36	46	57	68	78	87
Racking Costs	**$32,250**	**$41,493**	**$51,396**	**$61,046**	**$70,580**	**$78,266**
Total Network Costs	**$1,294,136**	**$264,417**	**$88,843**	**$67,407**	**$71,552**	**$78,519**
Number of CSRs	51	66	82	97	113	125
VoD Staffing Costs per CSR	$16,000	$16,544	$17,106	$17,688	$18,289	$18,910.00
Total Staffing Costs	**$823,760**	**$1,095,875**	**$1,403,547**	**$1,723,784**	**$2,060,713**	**$2,362,710**
Equipment Maintenance	$71,667	$46,068	$28,510	$16,918	$9,773	$5,414
BadDebt (@ 2% of Revenue)	$964,557.06	$1,178,783.72	$1,383,082.06	$1,596,998.98	$1,740,599.79	$2,340,809.59
Marketing/Promotion	$482,278.53	$589,391.86	$691,541.03	$798,499.49	$870,299.90	$5,852,023.98
Total Operating Costs	**$4,191,613.13**	**$4,351,804.41**	**$5,486,212.38**	**$6,495,893.87**	**$7,436,248.79**	**$13,533,442.09**
P&L						
Balance Sheet	**$7,846,059**	**$10,359,416**	**$11,774,652**	**$13,434,653**	**$14,286,437**	**$15,679,862**

[1]Boldface type indicates results while roman type indicates intermediate values.

work item to the whole international standards organization. Once the work item has been approved, working groups of experts define the scope, then proceed to negotiate the technical details of the specification. In the final phase, the voting members put the standard specification to a vote. If successful, the specification is published as a new standard. The standards are revised periodically. The simplest approach to verification of the implementation of standards is through test models and compliance bit streams.

Intellectual Property Issues

Although a standard may be defined and published, practicing that standard may require licensing a number of patents essential to performance of the standard. The MPEG LA organization (which grew out of the MPEG Intellectual Property Rights Group) maintains a list of about 20 patent holders who have decided to pool their patents for common licensing, a practice which has raised questions of antitrust law violation. Although the licensor MPEG LA did not collect royalties on MPEG-1, it did

Table 6 Hypothetical iMode™ Live Customer Service Business Model

Digital Video Business Model						
iMode Live Customer Service	**Year 1**	**Year 2**	**Year 3**	**Year 4**	**Year 5**	**Year 6**
Avg Subs	27,000,000	41,600,000	56,200,000	70,800,000	85,400,000	100,000,000
% who use customer service	**1.0%**	**1.0%**	**1.0%**	**1.0%**	**1.0%**	**1.0%**
Avg users per day	270,000	416,000	562,000	708,000	854,000	1,000,000
Conversion Rate (daily)	**1.0%**	**1.0%**	**1.0%**	**1.0%**	**1.0%**	**1.0%**
purchases/day	2,700	4,160	5,620	7,080	8,540	10,000
Average Purchase	**$20.00**	**$20.00**	**$20.00**	**$20.00**	**$20.00**	**$20.00**
Monthly Subscription per Sub	**$10**	**$10**	**$10**	**$10**	**$10**	**$10**
Gross purchase revenue per day	$54,000.00	$83,200.00	$112,400.00	$141,600.00	$170,800.00	$200,000.00
Gross subscription revenue per day	$90,000.00	$138,666.67	$187,333.33	$236,000.00	$284,666.67	$333,333.33
Gross revenue per day	$144,000.00	$221,866.67	$299,733.33	$377,600.00	$455,466.67	$533,333.33
days in year	**365**	**365**	**365**	**365**	**365**	**365**
Annual Gross Revenue	$52,560,000	$80,981,333	$109,402,667	$137,824,000	$166,245,333	$ 194,666,667
Content Partner Revenue Share	**90%**	**90%**	**90%**	**90%**	**90%**	**90%**
Content Partner Net Revenue	**$47,304,000**	**$72,883,200**	**$98,462,400**	**$124,041,600**	**$149,620,800**	**$175,200,000**
Carrier Revenue Share	**10%**	**10%**	**10%**	**10%**	**10%**	**10%**
Carrier Net Revenue	**$4,730,400**	**$7,288,320**	**$9,846,240**	**$12,404,160**	**$ 14,962,080**	**$17,520,000**
Capital Costs						
Maximum Concurrent Usage	**10%**	**10%**	**10%**	**10%**	**10%**	**10%**
New Streams Required	27000	41600	56200	70800	85400	100000.00
Server Capacity (bps)	200,000,000	200,000,000	200,000,000	200,000,000	200,000,000	200,000,000
Bandwidth per stream (kbps)	2	2	2	2	2	2
Number of New Servers	1	0	0	0	0	0
Cost per Server	$10,000.00	$4,996.17	$2,496.17	$1,247.13	$623.09	$311.31
Cost per Stream	$50.00	$50.00	$50.00	$50.00	$50.00	$50.00
New Server Costs	$10,000.00	$0.00	$0.00	$0.00	$0.00	$0.00
New Stream Costs	$1,350,000.00	$2,080,000.00	$2,810,000.00	$3,540,000.00	$4,270,000.00	$5,000,000.00
Total Server Cost	$1,360,000.00	$2,080,000.00	$2,810,000.00	$3,540,000.00	$4,270,000.00	$5,000,000.00
Amortized Server Costs	**$452,000.00**	**$1,145,333.33**	**$2,082,000.00**	**$2,812,000.00**	**$3,542,000.00**	**$4,270,000.00**
Operating Costs						
Bandwidth Costs						
Sessions per year	985,500	1,518,400	2,051,300	2,584,200	3,117,100	3,650,000
Average 5 min call (128-byte packets)	600	600	600	600	600	4176
Total Usage (packets)	591,300,000	911,040,000	1,230,780,000	1,550,520,000	1,870,260,000	15,242,400,000
Bandwidth costs ($ per packet)	3.E-02	4.E-03	6.E-04	9.E-05	1.E-05	2.E-06
Total Bandwidth Costs	**$15,373,800**	**$3,613,811**	**$744,840**	**$143,158**	**$26,345**	**$32,757**
Cost per Rack Unit	$900	$900	$900	$900	$900	**$900**
Rack Units (2RU servers)	1	0	0	0	0	0
Racking Costs	**$450**	**$0**	**$0**	**$0**	**$0**	**$0**
Total Network Costs	**$15,374,250**	**$3,613,811**	**$744,840**	**$143,158**	**$26,345**	**$32,757**
Number of CSRs	135	208	281	354	427	500
Vod Staffing Costs per CSR	$16,000	$16,544	$17,106	$17,688	$18,289	$18,910.00
Total Staffing Costs	**$2,160,000**	**$3,441,152**	**$4,806,786**	**$6,261,552**	**$7,809,403**	**$9,455,000**
Equipment Maintenance	$1,000	$0	$0	$0	$0	$0
BadDebt (@ 2% of Revenue)	$1,051,200.00	$1,619,626.67	$2,188,053.33	$2,756,480.00	$3,324,906.67	$3,893,333.33
Marketing/Promotion	$525,600.00	$809,813.33	$1,094,026.67	$1,378,240.00	$1,662,453.33	$9,733,333.33
Total Operating Costs	**$19,564,050.00**	**$10,629,736.69**	**$10,915,706.27**	**$13,351,429.72**	**$16,365,107.75**	**$27,384,423.37**
P&L						
Balance Sheet	**$27,739,950**	**$62,253,463**	**$87,546,694**	**$110,690,170**	**$133,255,692**	**$147,815,577**

on MPEG-2 and has proposed a licensing scheme for MPEG-4 that includes an hourly usage fee that could amount to as much as $0.25 per hour per user. ITU-T maintains a patent database (http://www.itu.int/ITU-T/dbase/patent/) and, in general, prospective developers must negotiate with patent holders for license terms.

DIGITAL VIDEO APPLICATION SOLUTIONS

Table 4 extends Table 1 by indicating additional well-known applications of video compression, together with requirements and possible solutions.

DIGITAL VIDEO BUSINESS MODELS

Video-on-Demand over Broadband Networks

In this application, customers request movies delivered via cable modem or digital subscriber line service to their personal computer. Brief inspection of the Intertainer Web site (http://www.intertainer.com) shows that full screen (800 × 600 pixels), full motion movies encoded at about 500 kbps are currently delivered via the Windows Media 8 platform at 580 kbps or higher including overheads so that a 120-min movie causes almost 4 GB of data transfer. Table 5 shows a hypothetical business model for a movie-on-demand service delivered in the United States.

Table 7 Hypothetical FOMA Live Adult Service Business Model

Digital Video Business Model						
FOMA Live Adult Service	**Year 1**	**Year 2**	**Year 3**	**Year 4**	**Year 5**	**Year 6**
Avg Subs	55,000	235,000	415,000	595,000	775,000	955,000
% who use VOD service	**75.0%**	**75.0%**	**75.0%**	**75.0%**	**75.0%**	**75.0%**
Avg users per day	41,250	176,250	311,250	446,250	581,250	716,250
Conversion Rate (daily)	**0.8%**	**0.8%**	**0.8%**	**0.8%**	**0.8%**	**1.0%**
purchases/day	342	1,463	2,583	3,704	4,824	7,163
Average Purchase	**$25.00**	**$25.00**	**$25.00**	**$25.00**	**$25.00**	**$20.00**
Monthly Subscription per Sub	**$10**	**$10**	**$10**	**$10**	**$10**	**$10**
Gross purchase revenue per day	$8,559.38	$36,571.88	$64,584.38	$92,596.88	$120,609.38	$143,250.00
Gross subscription revenue per day	$13,750.00	$58,750.00	$103,750.00	$148,750.00	$193,750.00	$238,750.00
Gross revenue per day	$22,309.38	$95,321.88	$168,334.38	$241,346.88	$314,359.38	$382,000.00
days in year	**365**	**365**	**365**	**365**	**365**	**365**
Annual Gross Revenue	$8,142,922	$34,792,484	$61,442,047	$88,091,609	$114,741,172	$139,430,000
Content Partner Revenue Share	**90%**	**90%**	**90%**	**90%**	**90%**	**90%**
Content Partner Net Revenue	**$7,328,630**	**$31,313,236**	**$55,297,842**	**$79,282,448**	**$103,267,055**	**$125,487,000**
Carrier Revenue Share	**10%**	**10%**	**10%**	**10%**	**10%**	**10%**
Carrier Net Revenue	**$732,863**	**$3,131,324**	**$5,529,784**	**$7,928,245**	**$10,326,705**	**$12,548,700**
Capital Costs						
Maximum Concurrent Usage	**10%**	**10%**	**10%**	**10%**	**10%**	**10%**
New Streams Required	4125	17625	31125	44625	58125	71625
Server Capacity (bps)	200,000,000	200,000,000	200,000,000	200,000,000	200,000,000	200,000,000
Bandwidth per stream (kbps)	384	384	384	384	384	384
Number of New Servers	8	35	61	88	114	141
Cost per Server	$10,000.00	$4,996.17	$2,496.17	$1,247.13	$623.09	$311.31
Cost per Stream	$50.00	$50.00	$50.00	$50.00	$50.00	$50.00
New Server Costs	$10,000.00	$0.00	$0.00	$0.00	$0.00	$0.00
New Stream Costs	$206,250.00	$881,250.00	$1,556,250.00	$2,231,250.00	$2,906,250.00	$3,581,250.00
Total Server Costs	$216,250.00	$881,250.00	$1,556,250.00	$2,231,250.00	$2,906,250.00	$3,581,250.00
Amortized Server Costs	**$70,750.00**	**$364,500.00**	**$883,250.00**	**$1,558,250.00**	**$2,233,250.00**	**$2,906,250.00**
Operating Costs						
Bandwidth Costs						
Sessions per year	124,967	533,949	942,932	1,351,914	1,760,897	2,614,313
Average 5-min call (128-byte packets)	115200	115200	115200	115200	115200	4176
Total Usage (packets)	14,396,184,000	61,510,968,000	108,625,752,000	155,740,536,000	202,855,320,000	10,917,369,000
Bandwidth costs ($ per packet)	4.E-04	6.E-05	9.E-06	1.E-06	2.E-07	3.E-08
Total Bandwidth Costs	**$5,614,512**	**$3,659,922**	**$986,068**	**$215,690**	**$42,862**	**$352**
Cost per Rack Unit	$900	$900	$900	$900	$900	**$900**
Rack Units (2RU servers)	4	17	31	44	57	70
Racking Costs	**$3,650**	**$15,593**	**$27,537**	**$39,481**	**$51,425**	**$63,369**
Total Network Costs	**$5,618,161**	**$3,675,516**	**$1,013,605**	**$255,172**	**$94,287**	**$63,721**
Number of CSRs	1	1	2	3	4	5
Total Staffing Costs	**$3,664,315**	**$15,656,618**	**$27,648,921**	**$39,641,224**	**$51,633,527**	**$62,743,500**
Equipment Maintenance	$1,000	$0	$0	$0	$0	$0
BadDebt (@ 2% of Revenue)	$162,858.44	$695,849.69	$1,228,840.94	$1,761,832.19	$2,294,823.44	$2,788,600.00
Marketing/Promotion	$81,429.22	$347,924.84	$614,420.47	$880,916.09	$1,147,411.72	$6,971,500.00
Total Operating Costs	**$9,598,513.80**	**$20,740,408.17**	**$31,389,037.70**	**$44,097,394.00**	**$57,403,299.52**	**$75,473,571.15**
P&L						
Balance Sheet	**−$2,269,884**	**$10,572,828**	**$23,908,804**	**$35,185,054**	**$45,863,755**	**$50,013,429**

The number of subscribers is based on the projected number of broadband users from Lathen (1999); the usage of video on demand (VoD) is projected from the Broadwing Investor Relations (1999) report on the usage of the Cincinnati Bell Zoomtown movie-on-demand system; the buy rates and studio paybacks are based on the most conservative model described in *Video on Demand 2001* (2001).

Projected server and stream costs are those indicated by Real Media (http://www.real.com) combined with Moore's Law (costs halve in 18 months) with software costs amortized by straight line depreciation over three years and hardware costs depreciated over five years; bandwidth and racking costs are from the New York Band-X exchange with double Moore's Law (costs halve in 9 months); staffing profile of 1 customer service representative (CSR) per 200,000 customers with burdened labor rate similar to the minimum wage in the United States and relatively conservative marketing and promotional costs of 1% gross revenue.

Customer Relationship Management over Third Generation Mobile Networks

As the saying goes, "People prefer to buy from people they trust." Many people trust information presented visually. Video-enabled commerce, developed independently by Avaya (http://www.avaya.com) and Media Logic Systems Ltd. (http://www.iseetv.net) can facilitates sales, develop customer relations, and prevent churn.

Lucent Technologies (1996) announced a multimedia call center which makes skilled employees available to customers at remote computers and kiosks via H.320 and T.120 video conferencing links over telephone lines. According to Lucent "video infomercials can educate or market the caller while in queue or on hold."

Table 6 shows a hypothetical business model for a live video customer service application in Japan using MPEG-4 facial animation at 2000 bps (Tekalp and Ostermann, not dated) delivered over NTT DoCoMo iMode screen phones with 9.6 kbps connectivity. Bandwidth costs and revenue share are based on the current iMode pricing model together with double Moore's Law cost reduction. Subscriber numbers are based on a current figure of 27,000,000 adding 40,000 subscribers per day (from http://www.imode.com).

Table 7 shows a hypothetical business model for a live video adult services application in Japan using ITU-T H.261 over H.324 terminals at 384 kbps delivered over the freedom of mobile multimedia access (FOMA) packet-switched system. Bandwidth costs and revenue share are based on the current FOMA pricing model (Imazu & Kuroda, 2001) together with double Moore's Law cost reduction. Subscriber numbers are based on a current figure of 55,000 (Ovum Research, 2002) adding 15,000 subscribers per month.

CONCLUSION AND FUTURE OUTLOOK

Three significant business opportunities have been presented with hypothetical business models together with areas for future progress in video compression research. It may be thought that the advent of high-performance high-bandwidth connection such as Internet2 would obviate the need for video compression. However, we have seen that increased availability leads to increased demand for bandwidth with applications such as remote telesurgery leading the way. The primary risks in the video compression field relate to ongoing patent litigation. Since "people prefer to buy from people" and many people trust information presented visually, video compression facilitates commerce by providing a significant component of infrastructure for visual electronic communications. More research is need in the areas of one and two-way mobile videoconferencing together with significant improvements of compression methods and more liberal licensing agreements.

GLOSSARY

Frame Image from video sequence.
Hyperspectral Recorded in multiple spectral bands.
Lossless Reversible, exact.
Lossy Irreversible, approximate.
Scene Sequence of frames.

CROSS REFERENCES

See *Data Compression; Speech and Audio Compression.*

REFERENCES

Aho, J. (2002, April 13). *A quick guide to digital video resolution and aspect ratio Conversions*. Retrieved May 8, 2002, from http://www.iki.fi/znark/video/conversion

American National Standards Institute (1996). *Digital transport of one-way video signals-parameters for objective performance assessment* [ANSI T1.803.03–1996]. Retrieved May 11, 2002, from American National Standards Institute Web site, http://www.ansi.org

Bell, S., Diaz, B., Holroyd, F., & Jackson, M. (1983). Spatially referenced methods of processing raster and vector data. *Image and Vision Computing, 1*(4), 211–220.

Bormans, J., & Hill, K. (2000). *MPEG-21: Defining and standardising a multimedia framework*. Retrieved August 15, 2002, from MPEG Web site, http://mpeg.telecomitalialab.com/documents/ibc2000_tutorial/Bormans_files/frame.htm

Broadband Week (2002, March 1). *Broadband direct* [NTT Docomo sees 6 Million 3G Subscribers by 2004]. Retrieved May 10, 2002, from Broadband Week Web site, http://www.broadbandweek.com

Broadwing Investor Relations. (1999, October 22). *Cincinnati bell delivers strong revenue and margin growth* [press release]. Retrieved May 12, 2002, from Broadwing Investor Relations Web site, http://investor.broadwing.com/news/19991022–13188.cfm

Cheung, K., & Tong, K. (1993, April 2). Proposed data compression schemes for the Galileo S-Band contingency mission. Paper presented at the 1993 Space and Earth Science Data Compression Workshop, Snowbird, Utah.

Chiariglione, L. (1996). *Coding of moving pictures and associated audio for digital storage media at up to about 1, 5 Mbit/s* [Short MPEG-1 description]. Retrieved

May 12, 2002, from MPEG Web site, http://mpeg.telecomitalialab.com/standards/mpeg-1/mpeg-1.htm

Colmenarez, A., Frey, B., & Huang, T. A. (1999, October 25–28). Detection and tracking of faces and facial features. Paper presented at the International Conference on Image Processing, Kobe, Japan.

Commission Internationale de l'Eclairage. (1991). *CIE standard colorimetric observers.* Retrieved May 10, 2002, from International Standardization Organization, http://www.iso.ch

Commission Internationale de l'Eclairage. (1999). *CIE standard illuminants for colorimetry.* Retrieved May 10, 2002, from International Standardization Organization, http://www.iso.ch

Commission Internationale de l'Eclairage. (2000). *CIE Web site of the International Commission on Illumination.* Retrieved May 10, 2002, from http://www.cie.co.at/cie/

Dawes, B. (1999, October 1). *The color metric converter.* Retrieved May 10, 2002, from http://www.colorpro.com/info/tools/convert.htm

Fisher, Y. (Ed.). (1995). *Fractal image compression: Theory and application.* New York: Springer-Verlag.

Freedman, I., Boggess, E., & Seiler, E. (1993). Systems aspects of COBE science data compression. In M. Cohn & J. Storer (Series Eds.), *Data Compression Conference: Data Compression Conference 1993* (p. 502). (From *NASA Space and Earth Science Data Compression Workshop: NASA CP 3191. 1993 Space and Earth Science Data Compression Workshop,* p. 35, by I. Freedman, E. Boggess, & E. Seiler, 1993, Washington DC: National Aeronautics and Space Administration).

Freedman, I., & Farrelle, P. M. (1996). Systems aspects of COBE science data compression. In G. H. Jacoby (Vol. Ed.), *Astronomical Society of the Pacific Conference Series: Vol. 101. Astronomical Data Analysis Software and Systems V* (pp. 72–75). San Francisco: Astronomical Society of the Pacific.

Gibson, L., & Lucas, D. (1982). Spatial data processing using generalized balanced ternary. *Proceedings of the IEEE Computer Society on Pattern Recognition and Image Processing,* 566–571. Los Alamitos, CA: IEEE Computer Society.

Grunbaum, B., & Sheperd, G. C. (1980). Tilings with congruent tiles. *Bulletin of the American Mathematical Society, 3*(3), 951–973.

Herring, C. (1994, May 1). *An architecture for cyberspace: Spatialization of the Internet.* Retrieved May 10, 2002, from Citeseer, http://citeseer.nj.nec.com/334929.html

Hunt, S. (2001, October 18). *Video and graphics distribution system for mobile users* [WO0177800, published patent application]. Retrieved May 12, 2002, from esp@cenet, http://l2.espacenet.com/espacenet/viewer?PN=GB2366148&CY=gb&LG=en&DB=EPD

Imazu, H., & Kuroda, R. (2001, June 1). (keitai-l) [Fwd: FOMA Data Card impression]. Message posted to KEITAI-L Web site electronic mailing list, archived at http://www.appelsiini.net/keitai-l/

International Standardization Organization (1993). *Information technology—Coding of moving pictures and associated audio for digital storage media at up to about 1, 5 Mbit/s—Part 2: Video* [ISO/IEc 11172–2:1993]. Retrieved May 12, 2002, from International Standardization Organization, http://www.iso.ch

International Telecommunication Union (1998, November). *ITU-R BT.470–6* [Conventional Television Systems, recommendation]. Retrieved May 8, 2002, from International Telecommunication Union, http://www.itu.ch

International Telecommunication Union (2000, March). *Methodology for the subjective assessment of the quality of television pictures.* Retrieved May 11, 2002, from International Telecommunication Union, http://www.itu.ch

Jackson, M., Bell, S., Stevens, A., Freedman, I., & Dickman, P. (1986). Efficiency of Tesseral arithmetic for 2.5-D image manipulation on serial computers and transputer arrays. In *Spatial Data Processing Using Tesseral Methods* (pp. 353–364).

Jain, A. K. (1989). *Fundamentals of digital image processing.* Englewood Cliffs, NJ: Prentice Hall International.

Kaup, A., & Panis, S. (1997, September). *On the performance of the shape adaptive DCT in object-based coding of motion compensated difference images.* Paper presented at 1997 Picture Coding Symposium, Berlin. Retrieved May 12, 2002, from Andre Kaup's Web page, http://www.lnt.de/~kaup/paper/pcs-97a.pdf

Koenen, R. (2001, December). *From MPEG-1 to MPEG-21: Creating an interoperable multimedia infrastructure* [ISO/IEC JTCA1/SC29/WG11 N4518]. Retrieved May 12, 2002, from MPEG Web site, http://mpeg.telecomitalialab.com/mpeg

Koenen, R. (2002, March). *MPEG-4 overview—(V.21_Jeju Version)* [ISO/IEC JTC1/SC29/WG11 N4668]. Retrieved May 12, 2002, from MPEG Web site, http://mpeg.telecomitalialab.com/standards/mpeg-4/mpeg-4.htm

Lathen, D. A. (1999, October). *Broadband today.* Retrieved May 10, 2002, from Federal Communications Commission Web site, http://www.fcc.gov

Lucent Technologies (1996, November 25). *Lucent Technologies announces new video multimedia call center.* Retrieved May 5, 2002, from Lucent TechnologiesWeb site, http://www.lucent.com/press/1196/961125.bca.html

Lucent Technologies (1996, November 25). *Lucent Technologies announces new video multimedia call center* [press release]. Retrieved March 27, 2002, from Lucent Technologies Web site, http://www.lucent.com/press/1196/961125.bca.html

Martinez, J. M. (Ed.). (2001, December). *Overview of the MPEG-7 standard (version 6.0)* [ISO/IEC JTC1/SC29/WG11 N4509]. Retrieved May 12, 2002, from MPEG Web site, http://mpeg.telecomitalialab.com/standards/mpeg-7/mpeg-7.htm

Morley, S. A. (2000, June 17). *Compression for digital cinema.* Paper presented at Infocomm, Anaheim CA. Retrieved May 9, 2002, from Qualcomm Digital Cinema-Technical Library, http://www.qualcomm.com/digitalcinema/pdf/compreso.pdf

National Aeronautic and Space Administration. (2001, August 20). *AVHRR land and LANDSAT images.* Retrieved May 11, 2002, from National Aeronautic and Space Administration Goddard Distributed Active Archive, http://daac.gsfc.nasa.gov

Nave, C. R. (2001). Light and vision. In *HyperPhysics*. Retrieved May 10, 2002, from HyperPhysics Web site, http://hyperphysics.phy-astr.gsu.edu/hbase/hframe.html

Ohanian, T. A. (1998). *Digital nonlinear editing*. Boston: Focal Press.

Ovum Research (2002). *FOMA: The first hundred days*. Retrieved May 12, 2003, from Ovum Research Web site, http://www.ovum.com

Pirazzi, C. *Lurker's guide to video*. Retrieved May 10, 2002, from lurkertech Web site, http://www.lurkertech.com

Poynton, C. (1999a, December 30). *Frequently asked questions about color*. Retrieved May 10, 2002, from Charles Poynton Web site, http://www.inforamp.net/~poynton

Poynton, C. (1999b, December 30). *Frequently asked questions about gamma*. Retrieved May 10, 2002, from Charles Poynton Web site, http://www.inforamp.net/~poynton

Riva, G., Gamberini, L., & Davide, F. (2001). Virtual reality in telemedicine. In G. Riva (Ed.), *Communications through virtual technology: Identity, community and technology in the Internet age* (pp. 101–114). Amsterdam: IOS Press.

Sarnoff Corporation (2001, June). *JND: A human vision system model for objective picture quality assessments*. Retrieved May 11, 2002, from Sarnoff Corporation Web site, http://www.sarnoff.com

Shannon, C. E. (1948). A mathematical theory of communication. *Bell Systems Technical Journal, 27,* 379–423, 623–656.

Shapiro, J. M. (1993). An embedded hierarchical image coder using zerotrees of wavelet coefficients. In J. A. Storer & M. Cohn (Series Eds.), *Data Compression Conference: Vol. DCC 1993* (pp. 214–223). Los Alamitos, CA: IEEE Computer Society Press.

Shi, Y. Q., & Sun, H. (2000). *Image and video compression for multimedia engineering: Fundamentals, algorithms and standards*. Boca Raton, FL: CRC Press.

Snyder, W. E., Qi, H., & Sander, W. (1999). A coordinate system for hexagonal pixels. *SPIE Medical Imaging 1999: Image Processing,* 716–727. Bellingham, WA: SPIE.

Tekalp, A. M. (1996). *Digital Video Processing*. Englewood Cliffs, NJ: Prentice Hall International.

Tekalp, A. M., & Ostermann, J. (n.d.). *Face and 3-D mesh animation in MPEG-4*. Retrieved May 12, 2002, from MPEG Synthetic/Natural Hybrid Coding Web site, http://leonardo.telecomitalialab.com/icjfiles/mpeg-4_si/8-SNHC_visual_paper/8-SNHC_visual_paper.htm

Teranex. (2002). *Correct cadence* [Application Note]. Retrieved August 15, 2002, from Teranex Web site, http://www.teranex.com/pdf/01-TER-152 Correct Cadence App Note.pdf

Video on demand 2001 (2001). Carmel, CA: Paul Kagan Associates, Inc.

Webster, A., Jones, C. T., Pinson, M. H., Voran, S. D., & Wolf, S. (1993, February). *An objective video quality assessment system based on human perception*. Paper presented at Human Vision, Visual Processing, and Digital Display IV, San Jose CA. Retrieved May 11, 2002, from CiteSeer, http://citeseer.nj.nec.com/webster93objective.html

Video Streaming

Herbert Tuttle, *The University of Kansas*

INTRODUCTION

It is difficult to describe and predict the growth of video streaming technology and applications. One may liken it to predicting the growth and integration of television in the early 1940s. For the sake of simplicity, a parallel will be drawn between television and video streaming technology and applications.

Streaming video, a relatively new technology, has made it easier for videos to be delivered in a digital format over computer networks, including the Internet and corporate networks, directly to personal computers. Streaming video involves taking video files, breaking them into smaller pieces, and sending them to their destination (RealNetworks, 2000). With a special application designed to handle video streams at the destination, users can play the video files as they are delivered or downloaded, instead of having to wait until the complete video is downloaded before they can play it. This eliminates the issues associated with having to download the entire video, including long wait times and disk space concerns.

Streaming video has given Internet Web site developers and administrators two primary options for presenting videos to their audiences. The first option is "video on demand," where the user can access the Web site and play the video at any time. Typically, the source material for video on demand is prerecorded and stored material. Most streaming applications on the Web use prerecorded material. The second option is "webcasting," where videos of specific live events are shown to many viewers at predetermined times. Webcasts are usually shown in real time; there is no prerecorded material. These events can include corporate meetings, news programs, concerts, and sporting events. These options make streaming video very attractive to viewers. With video on demand, a single video file can be accessed by hundreds of users. Webcasts are less expensive than satellite broadcasting, but they do require technical expertise to set up and deliver (DoIt & WISC, 2002).

Streaming video's utility is not limited to educational or commercial uses. With streaming video anyone can share his or her creation with others, such as an art work or a short film. Streaming video takes the everyday "cyberchat" to a more interactive level.

All indications are that the use of streaming video will continue to grow. The emergence of DSL- and cable-connected homes has provided a means to deliver near-broadcast-quality video to viewers (Microsoft.com, 2001a). Consumers want access to media throughout their homes and during their daily routines, such as while exercising or running errands. Companies are discovering the benefits of utilizing streaming video for training, "virtual" company meetings, and general communications. E-commerce businesses have found that streaming video can be another potential revenue source, with "pay per download" stream applications (Microsoft.com, 2001b). Artists have new audiences for their craft.

The streaming video technology is still evolving and has presented many opportunities and challenges to both developers and users.

BACKGROUND

The streaming video technology has been shaped by other previous major technological innovations such as television, movies, videos, personal computers, and modems. Other significant events that have impacted the development and use of streaming video are increased personal computer ownership, the development and fast-growing use of the Internet, and nationwide broadband deployment (Tanaka, 2000).

Of the technological innovations mentioned above, television has laid much of the groundwork upon which video and streaming video technologies are based. Many people were involved in the invention of television and it developed over several decades. The first television sets were actually mechanical in nature. However, the development of the cathode ray tube and amplifier tube provided the starting point for electronic television (Fortner, 2002). In 1923, Vladimir Kosma Zworykin, known as the father of modern television, invented the iconoscope. (Inventors Online Museum, 2002). The iconoscope allowed pictures to be electronically broken down into hundreds of thousands of elements. In 1932, Zworkin was

able to use the iconoscope to imitate the ways that human eyes view images for television broadcast (Inventors Online Museum, 2002). This technology was a key component in the advancement of electronic television and serves as the foundation for the design of the modern electronic televisions in use today (Fortner, 2002).

Although television may have provided the foundation for the technology of streaming video, the Internet has provided the means that has made it available to consumers in their homes and to businesses. The Internet has revolutionized the computer and communications world as never before. It has become a worldwide medium for broadcasting, information dissemination, collaboration, and interaction between individuals without regard to location (Leiner et al., 2000).

NETWORKING CONCEPTS

Because streaming video is delivered to the user over a network, it is important to understand the basics of how the information is handled and transmitted through a network. In essence, networking involves one computer exchanging information with another computer. Most Internet address begins with http://. HTTP stands for hypertext transfer protocol and is a standard or protocol (RealNetworks, 2000). It tells a browser and computer that HTML has been sent to it so it can read the incoming information.

In the case of some streaming video locations on the Internet, the addresses start with PNM://, RTP://, or RTSP://. PNM stands for progressive networks media and it is an older protocol. However, there are still a number of video clips in use that use this protocol (RealNetworks, 2000). RTP stands for real-time protocol, and it is one of the most commonly used protocols for streaming media on the Internet (Compaq Computer Corporation [Compaq], 1998). RTSP stands for real-time streaming protocol, which is the newest protocol (RealNetworks, 2000). In all three cases, these addresses tell a browser and computer that streaming video has been sent to it. It should be noted that any computer receiving streaming video must have a special application installed that can read and play the video. This topic will be discussed in more detail in the next section.

HOW STREAMING WORKS

Streaming involves taking video or audio files, breaking them down into packets of information, and sending them to their destination. At the receiving end, the viewer can then play the video as it is being downloaded. Because of the way that information flows on the network, it is easy to see that there would be a number of interruptions and delays in playing the video. To address this issue, a technology, called buffering, was developed to ensure that the playing of the video on the receiving end is smooth. Buffering is the process where a large number of information packets are collected before the playing of the video begins. Once enough packets have been collected, the playing of the video will begin. As the video plays, the buffering will continue until all of the information has been received. It is important to note that the video is not stored on the user's computer; it is received, buffered, and played.

The process described above is referred to as true streaming. It should not be confused with a method called pseudo-streaming or progressive download. Pseudo-streaming users wait until a significant portion of video file has been downloaded to their computer before viewing the video. This method allows users to save files to the hard drives on their computer for later viewing. Progressive download works best with very short media clips and a small number of simultaneous users (DoIt & WISC, 2002).

Streaming video may involve a video with or without sound. In the case of a video with sound, the visual portion of the video is delivered on one stream while the audio is delivered on another stream. Technology has been developed to synchronize these streams at the destination to ensure that the sound matches up with the action being viewed. Streaming files that include more than one medium are known as rich media. It should be noted that streaming can include slide presentations, text, video, audio, or any combination of these.

A number of components are required in order to make streaming video work on the Internet. First, the user must have a computer connection to the Internet via a local area network or modem. The user must also have a Web browser with the appropriate video player or plug-in installed. Many plug-ins can be downloaded from the Web for free. A plug-in works in conjunction with a browser to play streaming video files. A Web server stores Web pages or HTML files. Streaming video files are usually kept on a separate dedicated streaming server. When a streaming video link is clicked on a Web page, the browser reads the HTML code and the lets the player/plug-in take over (DoIt & WISC, 2002). The player accesses the selected video on the streaming server using the video protocols (RTP and RTSP) discussed in the previous section. After a few seconds of buffering, the video will start and play.

STREAMING TECHNOLOGIES AND SYSTEMS

A number of technologies are available for streaming video. The three major technologies are RealOne, QuickTime, and Windows Media (DoIt & WISC, 2002). Each streaming technology has three common hardware and/or software components: (1) servers and video files; (2) video players and plug-ins; and (3) compression, encoding, and creation tools (DoIt & WISC, 2002). The specifics of each technology will be discussed in more depth in a later section in this paper.

Each streaming technology mentioned above may have its own proprietary server and media file types that they use. Also, RealOne, QuickTime, and Windows Media have their own servers that stream files in their own formats. Therefore, it is important to create video files in a format that are compatible with the technology and server that will be used to stream the files. However, and relatively newer product called Helix offers open, comprhensive didital media communication for all players.

In order to play the video file, the user must have the second component, the player, installed on their computer. Users can download the player from the Web for free or, sometimes, it is included with the browser. As

with the video files and servers, each technology may have its own proprietary player. In some cases, one technology's media files cannot be played by another technology's player.

As indicated above, the third common streaming technology component is file creation, compression, and encoding. It involves the process of creating video files for streaming. Again, each technology may have its own proprietary way of creating, compressing, and encoding streaming video files. Therefore, special software may be needed to create streaming video files that are compatible with the video player on the receiving end.

The above discussion has focused on the system requirements for streaming video. At this point, it is worth noting that the typical streaming video system has five basic functions. First, the video must be captured, digitized, and then stored on a disk. Second, after the video is stored on a disk, it can be edited to improve its quality and content. Third, the video file must be compressed and encoded to the appropriate streaming format. Fourth, the video is delivered to the user via the video server. And, fifth, the user receives, decodes, buffers, and plays the video on the computer.

CAPTURING AND DIGITIZING VIDEO

In working with streaming video, the first step is to record the video or obtain a recorded video. There are two types of video that can be recorded. The first is analog video, which is produced with a vhs, hi-8, or beta cam format. The second is digital video, which is produced with a digital recorder or camera (DoIt & WISC, 2002).

Analog video contains video information in frames consisting of varying analog voltage values. It tends to degrade over time and it can contain imperfections such as snow in the picture. Digital video contains video information in a series of digital numbers that can be stored and transmitted without imperfections. Digital video does not degrade over time. The recent advances in the digital technology make it easier to store, retrieve, and edit digital video (Compaq, 1998).

If the video is from an analog source, it will have to be converted and compressed into a digital format. In order to do this conversion, an analog video capture card and the appropriate software will have to be installed on the computer. The video capture card is an expansion card that works in conjunction with, or replaces, the graphics adapter inside the computer. If the video is digital, a FireWire capture card can be used and the analog-to-digital step is not needed (Videomaker Magazine, 2001).

A side note on the digital video format that is worthwhile to review is that digital video often uses a different color format than the format used for computer monitors. Computer monitors display the color information for each pixel on the screen using the RGB (red, green, blue) format. (Pixels can be defined as the small elements or points that make up the frame.) Digital video frequently uses a format known as YCrCb, where Y represents the brightness (or luma) of a pixel, and Cr and Cb represent the pure color. In the different color schemes used in digital video, each pixel will have a brightness component but groups of pixels may share the CrCb color data. Hence, the terms 24-bit, 16-bit, and 12-bit color schemes refer to the number of color bits required per pixel (Compaq, 1998).

With the capture and conversion of the video, the video is transferred into a format that can be edited and then encoded for streaming. A number of formats are available. One of the most common of these is the AVI format. AVI stands for Audio Video Interlaced and was created by Microsoft. It is one of the oldest formats in use and is included with Microsoft's Windows applications (Fischer & Schroeder, 1996). This format was used in many of the early video editing systems and software. However, there are restrictions in using this format; the most notable of these is compatibility issues with some of the more advanced editing systems. Even with these issues, many editing systems and software can still use this format.

Another format is the MOV format, which was originally developed for the Macintosh computer by Apple. It then became the proprietary standard of Apple's QuickTime streaming technology (Fischer & Schroeder, 1996).

One of the most recent formats is the MPEG format. MPEG is a newer format and it is becoming very popular with streaming video users. MPEG stands for Motion Pictures Experts Group, which is an international organization that developed standards for the encoding of moving images (Fischer & Schroeder, 1996). There are a number of MPEG standards available, primarily for the encoding and compression of streaming video. These will be discussed in more detail later in this paper. However, one of the initial standards that was developed, MPEG-1, is used for the storage of video.

In capturing and converting video for streaming, it is recommended to maintain the highest quality video possible. The result will be very large video files that will have to be edited and streamed. However, it is better to start with the highest quality that can be maintained and then scale down to the quality that can be streamed. Starting with a lower quality leaves fewer options for editing, compression, and encoding.

EDITING THE VIDEO

Once the video has been captured and converted to a digital format, it can be edited with a variety of editing tools. As mentioned above, each of the three main streaming technologies—RealOne, QuickTime, and Windows Media—has editing tools. Editing is critical as it impacts how the video is ultimately received by the user and the end user's needs are paramount (see Producing Streaming Video for more.)

In editing a video, one of the first things that may have to be done is cropping the video. Cropping the video involves removing the edges, where electronic errors, glitches, and black bars may be seen. These usually appear during the process of recording and converting the video. In most cases, removing about 5% of the edges will eliminate the glitches. In cropping a video, it is important to remember that the final dimensions of the video must be compatible with the encoding technology (Kennedy, 2001).

Television systems use a technique called interlacing to display a picture on the screen. This process involves displaying the picture on every other line on the television

screen. Then lines are inserted between the first set. This process of alternating the picture lines eliminates any flicker on the screen. Videos also have this feature. However, for streaming video that will be displayed on a computer screen, interlacing is not needed. Some capture cards have a deinterlacing feature and some camcorders will record video without interlacing. However, if the video is interlaced at the editing step, and the file is very large, it is advisable to deinterlace the video during editing (Kennedy, 2001).

Also, when film is converted to video, additional frames are added because film is shot at 24 frames per second and depending on the video standard, television may run from 25 to 30 frames per second (Kennedy, 2001). The process of converting film to video, where the additional frames are put in, is called Telecine. It is best to avoid adding frames that are not needed. Therefore, if it is available, an Inverse Telecine conversion should be used to reduce the video back to 24 frames per second (Kennedy, 2001).

If a video has been shot with a lot of motion, the video could appear to be shaky or fuzzy, and not ideal for streaming. If this is the case, the best option may to use a still frame or slow motion. A still frame or slow motion may not look very natural, but it is better than streamed video that is not viewable.

Although special effects are great when viewed in a movie, they do not work well in streaming video because they utilize a lot of memory and impact the quality of the video. It is generally recommended that special effects be removed from the video. Streaming video is limited in its ability to deliver smooth video for any motion such as dance that relies on fluid movements. Also, if text is used in the video, it should be concise, legible, and easy to read.

Audio is a very important part of streaming video. If the video has an audio portion, the quality of the audio needs to be reviewed. For example, it is advisable to avoid the use of background music or other noise in order to ensure that speakers can be heard clearly. It is also good to prepare the audio to work on the worst speaker system that any potential user may have. If the audio is not clear then the usefulness of the video is greatly diminished.

BANDWIDTH

Before covering the topic of compressing and encoding, it is essential to understand the concept of bandwidth. The reason is that bandwidth is a critical factor in the transmission and reception of streaming video. Bandwidth is, simply put, the amount of information that can pass through a particular point of the wire in a specific amount of time (RealNetworks, 2000). Network bandwidth can be compared to a water pipe and a file to a tank of water. If the pipe is very narrow, then it will take a long time for the water from the tank to flow through the pipe. If the pipe is larger, then it will take less time for the water to flow through (Microsoft.com, 2000). Therefore, the higher the bandwidth, the greater the amount of information that can flow through the network to the destination. At the destination, the speed of the modem or other device used to connect to the Internet determines the bandwidth of the stream that is received.

Table 1 Available Bandwidths

Technology	Throughout
Fast Ethernet	100 Mbps
Ethernet	10 Mbps
Cable Modem	8 Mbps
ADSL	6 Mbps
1 × CD-ROM	1.2 Mbps
Single Channel ISDN	64 Kbps
High Speed Modem	56 Kbps
Standard Modem	28 Kbps

Because video files are large and many networks have limited bandwidths, there are many issues involved in transmitting these files over networks. Although many computer networks have installed new devices and technology to improve their bandwidths, this is one of the biggest challenges to streaming a video over a network. The Internet was not designed to handle streaming video.

File sizes are measured in kilobytes (abbreviated as K or KB). A kilobyte contains 1,024 bytes. When this conversion is applied to large video files and the math is done to determine transmission rates, it is apparent that these files have a huge amount of information that has to be transmitted. For example, a full-screen, full-motion video can require a data transmission rate of up to 216 Megabits per second (Mbps) (Compaq, 1998). This exceeds the highest available data rates in most networks. Table 1 shows the available bandwidth for several methods of data delivery, according to Compaq (1998).

In reviewing the above exhibit, it should be noted that the throughput listed for each technology represents an upper limit for that technology. In most cases, the actual throughput will be below this limit due to the amount of traffic on the network. Depending on the conditions of their connections, many users will see their data fluctuate up and down. One minute, they may have a 10 Kbps rate; the next minute, it may jump to 24 Kbps (Kennedy, 2000). Therefore, it is important for the provider of the streaming video to match the data rate to the conditions and limitations of the potential users.

Also, the Fast Ethernet and Ethernet technologies listed in Table 1 are used primarily in businesses and organizations. Single channel ISDN (integrated services digital network) is also used by businesses for video phones and video conferencing. Cable modems and ASDL (asymmetrical digital subscriber loops) are available to individual Internet users, but they are newer, more expensive technologies and are not as widely available as modems. Thus, it is safe to say that most Internet users have either a 56-Kbps high-speed modem or a 28-Kbps standard modem.

Two options are available for successfully delivering streaming video over networks. The first option involves scaling the video to smaller window sizes. This is important for low-bandwidth networks where many clients have modem access. The second option involves compressing the video using compression algorithms designed for this purpose. This is needed for most networks because of the

high bandwidth requirements of videos that have not been compressed.

Scaling and compressing video do affect the quality of the video. The quality of the video is impacted by frame rate, color, and resolution. Frame rate is the number of still images that make up one second of a moving video image. Images move fluidly and naturally at 30 frames per second, which is the National Television Standards Committee (NTSC) standard for full motion video. However, film is usually 24 frames per second (Compaq, 1998l). Videos with a frame rate of less than 15 frames per second become noticeably jumpy. It should be noted that most phone and modem technology limits the frame rate to 10 frames per second (Videomaker Magazine, 2001).

The second quality variable, color depth, is the number of bits of data the computer assigns to each pixel of the frame. The more bits of color data assigned to each pixel, the more colors can be displayed on the screen. Most videos are either 8-bit 256-color, 16-bit 64,000-color, or 24 bit 16.8-million color. The 8-bit color is very grainy and not suitable for video. The 24-bit color is the best, but it greatly increases the size of the streaming file, so the 16-bit color is normally used (Videomaker Magazine, 2001).

The third quality variable, resolution, is measured by the number of pixels contained in the frame. Each pixel displays the brightness and color information that it receives from the video signal. The more pixels in the frame, the higher the resolution. For example, if the video is 640×480, there are 640 pixels across each of the 480 vertical lines of pixels. Streamed video ranges from postage stamp size, which is 49×49 pixels, to full PC monitor screen, which is 640×480 pixels, and beyond (Videomaker Magazine, 2001).

SCALING

As mentioned previously, scaling involves reducing video to smaller windows. For example, this can be accomplished by reducing the frame resolution from a full screen (640×480) to a quarter screen (320×240). In addition, frame rate and color depth can also be scaled. For example, the frame rate can be reduced from 30 to 15 frames per second. The color depth can be scaled from 24-bit to 16-bit. According to Compaq (1998), the process noted in this example would reduce the video file size from 216 Mbps to 18 Mbps and the quality of the video would be reduced. However, as can be seen from the available bandwidths shown in Table 1, many delivery methods would not support a data rate of 18 Mbps. Therefore, to further reduce the data rate, video compression is necessary.

COMPRESSING AND ENCODING

The goal of compression is to represent video with as few bits as possible. Compression of video and audio involves the use of compression algorithms known as codecs. The term *codec* comes from the combination of the terms encoder and decoder—cod from encoder and dec from decoder (RealNetworks, 2000). An encoder converts a file into a format that can be streamed. This includes breaking a file down into data packets that can be sent and read as they are transmitted through the network. A decoder sorts, decodes, and reads the data packets as they are received at the destination. Files are compressed by encoder/decoder pairs for streaming over a network.

Encoders generally accept specific input file formats used in the capture and digitizing process. The encoders then convert the input formats into proprietary streaming formats for storage or transmission to the decoder. Some codecs may be process-intensive on the encode side in order to create programs one time that will be played many times by the users. Other codecs are divided more equally between encoding and decoding; these are typically used for live broadcasts (Compaq, 1998).

As mentioned above, each of the three major streaming technologies has its preferred encoding and compressing formats. Many users opt to work with one of these three technologies because they are relatively easy to use, and technical support is provided by each of the technologies. These technologies provide options to users for selecting video quality and data transmission rates during the compression and encoding process. Depending on the application and technology used, multiple streaming files may have to be produced to match the different bandwidths of the networks over which the video is streamed. Two of the three major technologies have advanced options where a streaming file can be produced that has a data transmission rate that will adapt to the varying bandwidths of the networks. The specifics of these technologies will be discussed in a later section.

Even with the dominance of the three major technologies, there are some open standards for compression algorithms. It is important to be aware of these standards and understand how the compression algorithms work. With this knowledge, the user can make better decisions when creating, delivering, and viewing streaming video. The compression algorithms will be discussed in more detail later. However, they all utilize the same basic compression techniques to one degree or another. Therefore, it is essential to review the compression techniques before discussing the algorithms.

First, compression techniques are either lossless or lossy. Lossless compression is a process where data are compressed without any alteration of the data in the compression process. There are situations where messages must be transmitted without any changes. In these cases, lossless compression can be used. For example, lossless compression is typically used on computers to compress large files before emailing them (Vantum Corporation, 2001). A number of lossless techniques are available. However, for video files in particular, more compression is needed than the lossless techniques can provide.

Lossy techniques involve altering or removing the data for efficient transmission. With these techniques, the original video can only be approximately reconstructed from its compressed representation. This is acceptable for video and audio applications as long the data alteration or removal is not too great. The amount of alternation or removal that is acceptable depends on the application (Vantum Corporation, 2001).

A number of video compression techniques take advantage of the fact that the information from frame to frame is essentially the same. For example, a video that shows a person's head while that person is talking will have the same background throughout the video. The only changes will be in the person's facial expressions and other

gestures. In this situation, the video information can be represented by a key frame along with delta frames containing the changes between the frames. This is known as interframe compression. In addition, individual frames may be compressed using lossy techniques. An example of this is a technique where the number of bits representing color information is reduced and some color information is lost. This is known as intraframe compression. Combining the interframe and intraframe compression techniques can result in up to a 200:1 compression (Compaq, 1998).

Another compression technique is called quantizing. It is the basis for most lossy compression algorithms. Essentially, it is a process where rounding of data is done to reduce the display precision. For the most part, the eye cannot detect these changes to the fine details (Fischer & Schroeder, 1996). An example of this type of compression is the intraframe compression described above. Another example is the conversion from the RGB color format used in computer monitors to the YcrCb format used in digital videos that was discussed in the capturing and digitizing section of this paper.

Filtering is a very common technique that involves the removal of unnecessary data. Transforming is another technique, where a mathematical function is used to convert the data into a code used for transmission. The transform can then be inverted to recover the data (Vantum Corporation, 2001).

For videos that have audio, the actual process used to compress audio is very different from that used to compress video even though the techniques that are used are very similar to those described above. This is because the eye and ear work very differently. The ear has a much higher dynamic range and resolution. The ear can pick out more details but it is slower than the eye (Filippini, 1997). Sound is recorded as voltage levels and it is sampled by the computer a number of times per second. The higher the sampling rate, the higher the quality and hence, the greater the need for compression. Compressing audio data involves removing the unneeded and redundant parts of the signal. In addition, the portions of the signal that cannot be heard are removed.

VIDEO COMPRESSION ALGORITHMS

Some algorithms were designed for wide bandwidths and some for narrow bandwidths. Some algorithms were developed specifically for CD-ROMs and others for streaming video. There are a number of compression algorithms available for streaming video; this chapter will discuss the major ones in use today. These algorithms are MPEG-1, MPEG-2, MPEG-4, H.261, H.263, and MJPEG. The video compression algorithms can be separated into two groups: those that make use of frame-to-frame redundancy and those that do not. The algorithms that make use of this redundancy can achieve significantly greater compression. However, more computational power is required to encode video where frame-to-frame redundancies are utilized.

As mentioned in earlier in this paper, MPEG stands for Moving Pictures Experts Group, which is a work group of the International Standards Organization (ISO) (Compaq, 1998). This group has defined several levels of standards for video and audio compression. The MPEG standard only specifies a data model for compression and, thus, it is an open, independent standard. MPEG is becoming very popular with streaming video creators and users.

The first of these standards, MPEG-1, was made available in 1993 and was aimed primarily at video conferencing, videophones, computer games, and first-generation CD-ROMs. It was designed for consumer video and CD-ROM audio applications that operate at a data rate of approximately 1.5 Mbps and a frame rate of 30 frames per second. It has a resolution of 360×242 and supports playback functions such as fast forward, reverse, and random access into the bitstream (Compaq, 1998). It is currently used for video CDs and it is a common format for video on the Internet when good quality is desired and when its bandwidth requirements can be supported (Vantum Corporation, 2001).

MPEG-1 uses interframe compression to remove redundant data between the frames, as discussed in the previous section on compression techniques. It also uses intraframe compression within an individual frame as described in the previous section. This compression algorithm generates three types of frames: I-frames, P-frames, and B-frames. I-frames do not reference other previous or future frames. They are stand-alone or Independent frames and they are larger than the other frames. They are compressed only with intraframe compression. They are the entry points for indexing or rewinding the video, because they represent complete pictures (Compaq, 1998).

On the other hand, P-frames contain predictive information with respect to the previous I or P frames. They contain only the pixels that have changed since the last frame, and they account for motion. In addition, they are smaller than the I-frames, because they are more compressed. I-frames are sent at regular intervals during transmission process. P-frames are sent at some time interval after the I-frames have been sent (this time interval will vary based on the transmission of the streaming video).

If the video has a lot of motion, the P-frames may not come fast enough to give the perception of smooth motion. Therefore, B-frames are inserted between the I- and P-frames. B-frames use data in the previous I- or P-frames as well as the future I- or P-frames, thus, they are considered bidirectional. The data that they contain are an interpolation of the data in the previous and future frames, with the assumption that the pixels will not drastically change between the two frames. As a result, the B-frames have the most compression and are the smallest of the three types of frames. In order for a decoder to decode the B-frames, it must have the I- and P-frames that they are based on; thus the frames may be transmitted out of order to reduce decoding delays (Comqaq, 1998).

A frame sequence consisting of an I-frame and its following B- and P-frames before the next I-frames is called a group of pictures (GOP) (Compaq, 1998). There are usually around 15 frames in a GOP. An example of the MPEG encoding process can be seen in Figure 1. The letters I, P, and B in the figure represent the I-, P-, and B-frames that could possibly be included in a group of pictures. The letters were sized to indicate the relative size of the frame (as compared to the other frames).

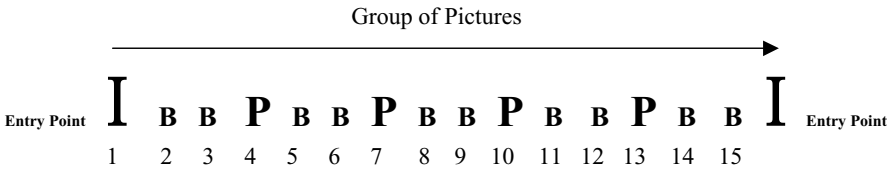


Figure 1: MPEG-1 encoding process.

One disadvantage of the MPEG format is that it cannot easily be edited because video cannot be entered at any point. And the quality of the resulting video is impacted by the amount of motion in the video. The more motion in the video, the greater the probability that the quality will be reduced. The MPEG encoding and decoding process can require a large amount of computational resources, which requires the use of specialized computer hardware or a computer with a powerful processor.

MPEG-2 was released in 1994 and was designed to be compatible with MPEG-1. It is used primarily for delivering digital cable and satellite video to homes. It is the basis of DVD and HDTV. MPEG-2 utilizes the same compression techniques as MPEG-1. However, it has been enhanced so that it has better compression efficiency than MPEG-1. MPEG-2 supports two encoding schemes depending on the application. The first scheme has a variable bit rate, which keeps the quality constant. The second scheme involves varying the quality to keep the bit rate constant. MPEG-2 is not considered an ideal format for streaming over the Internet because it works best at transmission rates higher than most networks can handle (Cunningham & Francis, 2001).

MPEG-4 is one of the most recent video formats and is geared toward Internet and mobile applications including video conferencing, video terminals, Internet video phones, wireless mobile video, and interactive home shopping. It was originally designed to support data rates less than 64 Kbps but has been enhanced to handle data rates ranging from 8 Kbps to 35 Mbps. MPEG-4 is different from MPEG-1 and -2 in that it has been enhanced to handle the transmission of objects described by shape, texture, and motion, versus just the transmission of rectangular frames of pixels. In fact, it is very similar to H.263, which is the video conferencing standard (Compaq, 1998). This feature makes MPEG-4 well suited to handle multimedia objects, which are used in interactive DVD, interactive Web pages, and animations.

MPEG-7 is the newest standard. It is designed for multimedia data and can be used independent of the other MPEG standards. Work is being done on an extension of the MPEG-7 standard, called MPEG-21.

The H.261 and H.263 standards are designed for video conferences and video phone applications that are transmitted over an ISDN network. H.261 has the ability to adapt the image quality to the bandwidth of the transmission line. The transmission rate for H.261 is usually around 64 Kbps (Fischer & Schroeder, 1996). H.263 was developed as an enhancement to H.261 and was designed to support lower bit rates than H.261. It has a higher precision for motion compensation than H.261. H.263 is very similar to the MPEG standards, particularly MPEG-4, and uses the same compression techniques (Vantum Corporation, 2001).

MJPEG stands for Motion JPEG, and JPEG stands for Joint Photographic Experts Group. JPEG is an international standard for compressing still frames. MJPEG is a sequence of JPEG compressed still images that represent a moving picture. Thus, MJPEG is a compression method that is applied to each frame without respect to the preceding or following image (Vantum Corporation, 2001). MJPEG can be edited easily but it is not able to handle audio.

AUDIO COMPRESSION ALGORITHMS

Each of the three major streaming technologies has its preferred algorithms for compressing audio. In addition, the MPEG group has defined an audio standard called MPEG-1 for audio. As discussed previously, audio compression is different than video, although it uses similar techniques. The MPEG audio compression uses psychoacoustic principles, which deal with the way the human brain perceives sound (Filippini, 1997).

The first principle utilized in the MPEG audio compression is the masking effect. This means that weak sounds are not heard, or they are masked, when they are near a strong sound. For example, when audio is digitized, some compression occurs because data are removed and noise is added to the audio. This noise can be heard during silent moments, or between words or sentences. However, this noise is not heard during talking or when music is playing. This is because the noise is a weaker sound and is masked by the louder talking or music. MPEG uses this masking effect to raise the noise floor around a strong sound because the noise will be masked anyway. And, by raising the noise floor, fewer data bits are used, and the signal (or file) is compressed. MPEG uses an algorithm to divide up the sound spectrum into subbands. It then calculates the optimum masking threshold for each band (Filippini, 1997).

The second psychoacoustic principle is that the human ear is less sensitive to high and low frequencies, versus middle frequencies. In essence, MPEG employs a filtering technique along with the masking effect to remove data from the high and low frequencies where the changes will not be noticed. It maintains the data in the middle frequencies to keep the audio quality as high as possible.

DELIVERING THE VIDEO

Once the video has been compressed and encoded for streaming, the next step is to serve the video to the users on the Internet. As discussed earlier in this chapter, delivering video over the Internet is usually accomplished with a streaming server, instead of a Web server. A streaming

server has some specialized software that allows it to manage a data stream as it is being transmitted through the network. It utilizes the streaming protocols (RTP and RTSP) to transmit the video file.

A Web server can be used to stream video, but it has been designed to transfer text and images over the Internet and it does not have the means to control a stream (Strom, 2001). When a Web server is used, a user selects a video file and it starts to be copied down to the PC using HTTP like any other data source on the Internet. The player takes control and the video is buffered and played. But because the Web servers are not able to control the stream, the delivery of the video can be erratic and the user could experience rebuffering interruptions. Thus, it is best to use a video server to ensure that the user will have a smooth playback without interruptions.

Video servers have capacity limitations, and they can only deliver a certain number of streams at any one time. The capacity of a server is measured in the number of simultaneous streams that it can put out at any given point in time. This can range from 20 to 5,000 or more, depending on the type of server (DoIt & WISC, 2002). If a user tries to access a video file after the server has reached its maximum capacity, the user will get a message stating that the server is busy and to try playing the video again after 1 or 2 minutes.

It is essential to note that streaming servers require the appropriate hardware, network connections, and technical expertise to set them up and administer them. This can consume time and resources, so many people and businesses choose to outsource this task to a host. A host is an agent or department that has the facilities and technical expertise to serve other people's streaming videos and other media content (DoIt & WISC, 2002). Hosts usually charge a fee for their services. There are numerous hosts that advertise on the Internet. When selecting a host, it is important to ensure that they can support the streaming technology being used by the client.

When using a host, the client will be able to transfer media files from his or her local computer to a streaming server. This is usually done by using special software called an FTP client (DoIt & WISC, 2002i). The host will set the person up with a password-protected account and a designated amount of server space. With this situation, the person may have text and graphics for a Web site residing on a Web server. Then, he or she has streaming files on a streaming server. This can be managed by using certain HTML tags on the Web page that will trigger and control the playback of the media files from the streaming server. This involves specifying the path of the particular video file on the streaming server. Each of the three major streaming technologies has its own unique embedded HTML tags for controlling the video files on servers. Many of the encoding applications can generate these HTML tags (DoIt & WISC, 2002).

As covered earlier in the discussion on bandwidth, not all networks are suited for the streaming of video. Video works best when the bandwidth of the network is continuously high. However, when the bandwidth of the video exceeds that of the network, delays in the transmission of the data packets can occur. These delays will cause the picture to flicker and the audio (if present) to start

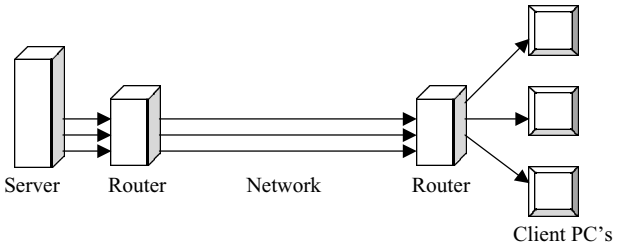


Figure 2: Video on demand.

and stop. In order to deal with the issues of streaming video and media, a new measure of network capability have been developed. It is called quality of service (QoS) (Compaq, 1998). Networks that have a good QoS measure provide a guaranteed bandwidth with few delays. The networks that have the best QoS are those that have dedicated connections for streaming.

Another network characteristic that needs to be considered for streaming video is the network's ability to support video-on-demand delivery and webcasting delivery. With video on demand (also know as unicasting), a stream is delivered onetoone to each client, and the user can request the video at any time. This type of delivery can consume a lot of network bandwidth, depending the number of users requesting a video. According to Compaq (1998), Figure 2 shows a simple diagram of how video on demand works. Each line in the exhibit represents a separate stream.

Webcasting, is used for live events where there can be potentially many viewers. Webcasting delivers one stream to many clients simultaneously. It does not consume as much bandwidth as video on demand. But as noted previously, video on demand is much more common because of the convenience it offers to users. Webcasting is scheduled for specific times and requires a lot of effort and resources to coordinate. Networks that support webcasting must have routers that are multicast capable. Figure 3 shows how a webcast works, according to Compaq (1998). The lines in the exhibit represent the video stream (note the single line going across the network).

RECEIVING, DECODING, AND PLAYING THE VIDEO

Finally, at the client desktop, the user accesses the video file. As discussed above, the user clicks on the video file that he or she wants to view, the request is routed to the appropriate file on the video server, and the player technology on the user's PC takes control of the data

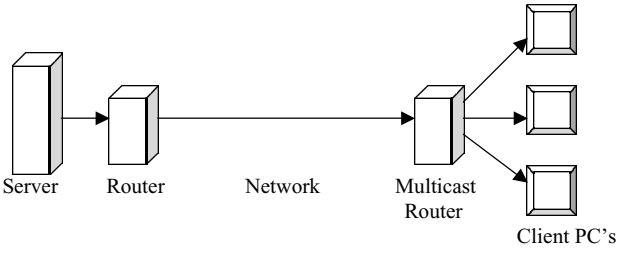


Figure 3: Webcasting.

transmission. The player buffers the stream(s), decodes the data packets, and converts the information back to analog so the video can be viewed. The player usually has functionality that will allow the user to play, pause, rewind, and fast forward. The player technology must be compatible with the streaming technology in order for the user to view the video.

With video on demand, the user can control access to the video. He or she can start, stop, rewind, and so forth at will. Although this freedom is desirable to the user, it does consume bandwidth on the network (as noted above). With webcasting, the user can only watch the video stream as it is being transmitted; he or she does not have any control over the stream. Webcasting does not use as much bandwidth as video on demand.

PRODUCING STREAMING VIDEO

Up to this point, this discussion has covered the aspects of streaming video after it has been created or produced. However, there are some techniques that should be used when used when producing streaming video that will make the capturing, editing, compressing, and encoding processes go much smoother. Because streaming video has to be compressed before delivery on the network, one of the most important things to remember when producing the video is to minimize motion and changes in the objects or people in the video. The more motion and change there is in the video, the more the video will have to be compressed and thus data (such as fine details and color) will be altered or removed.

Therefore, it is best to use a tripod or other method to anchor the camera whenever possible. If the camera is held in the video producer's hands, all of the hand movements will be incorporated into the video. The video producer should also avoid panning the camera as much as possible and avoid zooming in and out on a scene. Thus, eliminating the movement of the camera and keeping zooming in and out to a minimum will prevent changes from being introduced into the video.

The video producer should also try to keep the background as simple and consistent as possible. The producer should avoid trees, buildings, and so forth that will add complexity to the video, which will mean more data to compress. In addition, the producer should try to stay as close to the subject as possible when shooting the video. There may be some temptation to choose a wide shot of the scene however, will viewed online, the video will seem fuzzy. It is important to remember that the compression will remove a lot of the fine detail of the wide shot.

Last, the video producer should use an external microphone whenever possible. With an external microphone, the producer can keep it as close to the subject as possible to get good quality audio. With good quality audio, the audio compression will work much better. Audio is just as important as the images being displayed in the video.

VIDEO STREAMING USES

The previous sections have focused on the technical aspects of creating, delivering, and playing streaming video. This section will focus on the many uses of streaming video and the preferences of users. First, streaming media (video along with audio) have grown rapidly over the last few years. The number of Internet sites transmitting streaming video grew from 30,000 in mid-1998 to 400,000 by late 1999. The Net Aid concert in October, 1999, set a world record for the largest Internet broadcast event for a single day—2.5 million streams. The BBC Online's European solar eclipse site served a million streams in a day in August, 1999. The BBC estimated that its streaming audience was growing by 100% every 4 months (Tanaka, 2000).

As can been seen from the above statistics, streaming video continues to gain in popularity even with the technical challenges involved in streaming the video over the Internet. A Web survey conducted by Tanaka (2000) indicates that streaming video appeals to users because they can select what they want to view when they want to view it. Users like the fact that streaming technology has made specialized or unique videos or other or media available to them.

The streaming video uses fall into the primary categories of entertainment, news/information, education, training, and business. Entertainment was one of the earliest uses of streaming video and still remains the primary use of streaming video today. Entertainment covers a wide range of media including movies, music, and TV shows. There are numerous Web sites promoting free and pay-per-view movies. Many sites feature independent film makers, foreign films, and pornography (Bennett, 2002). Pornography video sites are some of the oldest entertainment sites on the Internet. At this time, pornography may be the largest online movie market of all on the Web (Bennett, 2002).

Recently, some Web sites have been established that show hit movies on the Internet. For example, viewers can watch the blockbuster movies on a Web site for $3.95 (Graham, 2002). The Hollywood studios have been slow to utilize the Internet as a medium for showing their movies because they want to be sure that this is a safe way to deliver their films. It is interesting to note that many movies are available on the Internet in unauthorized versions. Many of these movies were copied from DVDs or shot on a camcorder in a theater and then traded on file-sharing sites such as Morpheus and Kaaza (Graham, 2002).

According to Graham (2002), users need a high-speed Internet access, such as cable or ADSL, in order to watch movies over the Internet. Even with high-speed access, users may experience stutter, or stopping and starting, if there is a lot of traffic on the network. Users will be able to view the movie only on a partial or full PC screen size window.

Development in the streaming video world has been the integration of over-the-air and online entertainment programs. For example, in November, 1999, ABC.com and Warner Brothers Online hosted a simulcast of an episode of the Drew Carey Show. The television audience watched Drew's daily activities, while the Internet audience saw footage of what was happening in his home when he was out at work. ABC indicated that approximately 650,000 streams were served (Tanaka, 2000).

In the news/information category, many users like to utilize the Internet to view video clips of domestic news and international news items. Other users tend

to gravitate toward sites that provide clips on sporting events. For example, the on-air rating of the JFK Jr. tragedy was only 1.4, while over 2.3 million streams were delivered from the CNN.com Web site (Tanaka, 2000).

The use of streaming video in the education and training areas has been rapidly growing for the last few years. Universities and colleges, in particular, have been exploring the use of streaming video for their distance learning programs. Distance learning has become popular because many people who have been in the work force for a few years are returning to school to obtain an advanced degree, pursue a career change, or upgrade their skills. Many opt for distance learning programs because of their work or travel schedules, or because the academic programs they desire are not available locally. Because of the growth of distance learning programs, many colleges and universities have started to use streaming video as an alternative to mailing out VCR tapes, which can be cumbersome. With streaming video, these institutions can expand their distance learning programs to meet the needs of their students. In addition, there are many Web sites that offer training and tutorial programs on a variety of subjects.

A common presentation method used by educators is a lecture that includes static slides. These is must easier to create and will provide good quality sound and images for those students who have modem connections. There are a number of software tools that can be used to combine PowerPoint slides with narrations to create streaming presentations.

In view of the above discussion, it should also be pointed out that streaming video is used for teaching material that involves motion or dynamic interaction. Some examples of this include medical or laboratory procedures, processes in the physical sciences, interpersonal skills, and illustrations of real world events or activities (DoIt & WISC, 2002). In addition, live training or teaching webcasts are produced using audio, slides, or video. The participants access the Web site from their computers. Interaction between the instructor and participants occurs in real time. The participants can use a chat window to type in questions to the presenter during the session (DoIt & WISC, 2002). These events are very challenging to coordinate and deliver and are not as common as illustrated audio presentations.

Businesses and companies are starting to use streaming videos for advertising and communications. Some businesses have started to webcast their products in order to improve their sales. One of the most talked about events was the Victoria's Secret fashion show that was webcast in February, 1999 (Tanaka, 2000).

Another form of advertising that has become increasingly popular is the video banner ad (Tanaka, 2000). This technology involves using a program that detects whether or not the client PC has a streaming media player, and then determines the type if there is one present. This is done before the user clicks on the Web page. Once the user clicks on the Web page, video is played using the media player on the PC. If there is no media player on the PC, then a regular GIF banner is displayed.

Businesses are also using streaming media to broadcast presentations, corporate meetings, and in-house seminars to their employees. Many companies are finding that this is less expensive than live meetings and seminars, where travel expenses are incurred. And it offers opportunities for communication that would not otherwise be available. For example, a company that uses streaming technology may choose to broadcast an industry analysts' meeting or public relations event that, without this technology, would not be feasible to do.

THE BIG THREE STREAMING TECHNOLOGIES

As mentioned previously in this chapter, there are three major technologies for streaming video: RealOne, QuickTime, and Windows Media. These three players provide all of the tools needed for streaming video, including applications for creating, editing, compression, encoding, serving, and playing. Of these three, RealOne is the oldest and still the most widely used (Sauer, 2001). RealOne claims that they have over 70% of the Internet stream market with their player being installed in over 90% of home PCs (Cunningham & Francis, 2001). The RealOne technology supports over 40 media formats and employs the latest generation of encoding and compression techniques. They have also developed a technology, called *Surestream,* that utilizes an automatic bit-rate technology to adjust the data stream rate to the bandwidth characteristics of the user (Cunningham & Francis, 2001).

RealOne has developed some strategic partnerships that may give it a competitive advantage for the near future. First, RealOne now supports Apple's QuickTime technology. And it is working with the National Basketball Association and the Major Baseball League on a pay-per-view model (Cunningham & Francis, 2001). However, only the basic player and server versions are free; the more advanced server and productions tools available from RealOne can cost up to several thousand dollars. Streaming is ReadMedia's core business and they must charge fees for the use of their applications, whereas their competitors can incorporate their streaming technology into other products they sell, such as operating systems (Cunningham and Francis, 2001).

Also, RealOne is SMIL compliant. SMIL stands for synchronized multimedia integration language and it provides a time-based synchronized environment to stream audio, video, text, images, and animation (Strom, 2001). SMIL is a relatively new language available to streaming users. It is the officially recognized standard of the World Wide Web Consortium (Strom, 2001). SMIL has attracted a lot of attention because of the features and flexibility it offers to users.

QuickTime was developed by Apple in 1991 and it is one of the oldest formats for videos that are downloaded. It is the one of the recent entrants into the streaming video market (Sauer, 2001). One of the advantages that QuickTime offers is that it can support different compression techniques, including those used by RealOne, as noted above. QuickTime also features an open plug-in function that will allow the utilization of outside compression techniques (Cunningham & Francis, 2001). It is also SMIL compliant (as noted above for RealOne). Quicktime is available in the Apple MAC operating system. But it

offers a basic player and server tools that are compatible with other operating systems for free. It does charge a fee for the more advanced systems. Those advanced systems have been used by many in the cinematography field for editing purposes.

Windows Media was developed by Microsoft Corporation and it is a newcomer into the streaming video market. Since its introduction, it has been rapidly gaining ground on the other two technologies. Microsoft includes MediaPlayer as part of its Windows operating system, which can be a convenience for users. However, MediaPlayer is limited in its flexibility in that it has its own proprietary compression methods and it does not support many of the compression techniques utilized or developed by other companies. It does have a MPEG-4 type of compression algorithm and its proprietary compression methods are considered to very good. Microsoft does have the player and server tools available as free downloads from the Internet. And it has developed a technology called *Microsoft's Intelligent Streaming* that is like RealOne's *Surestream*. It allows the user to put multiple tracks, each with a different bit rate, into a single streaming file. This will allow the streaming file to adjust to fluctuations in the network's bandwidth (Cunningham & Francis, 2001).

OTHER STREAMING VIDEO SYSTEMS

The previous section covered the major technologies in the streaming video arena. However, a number of other key players have contributed to the growth of the streaming video field. The first of these is a company called Sorenson Media, which specializes in compression technologies. They are known as having the highest quality video compression, particularly for high motion at low data rates (Segal, 2002a). Sorenson Media also developed a professional version of their codec and a live broadcasting tool for Quicktime. And they have entered the hosting and streaming markets (Segal, 2002a). In fact, they were asked by the Church of Jesus Christ of Latter Day Saints to host and stream their semiannual conferences live and then archive the conferences for them (Segal, 2002a). And they worked with other companies, such as Macromedia, to build products that will incorporate streaming media.

Video hosting is another area that has several players. There are a number of companies that specialize in video hosting. Many offer services for each of the three major streaming technologies as well as other independent technologies. However, a potential user of video hosting would need to do some in-depth research before choosing a host.

In addition, there are companies that specialize in online broadcasting for TV news and programming. In fact, some of the major networks have their own online broadcasting Web sites. These include CNN, ABC, and BBC. In addition, some local TV networks in the larger metropolitan areas have their own Web sites. There are a few independent companies that specialize in online broadcasting. One that was reviewed was Servecast, which indicated that they would provide services for sporting events and other media. They also indicated that they could provide content protection.

For streaming video creation, editing, and encoding, a number of independent technologies are available. For example, Sonic Foundry has a tool that provides for the creation of streaming content in RealOne and Microsoft Media formats. And Terran has a tool called media cleaner that provides a complete set of tools for preparing video and audio for the Web. It is considered the industry leader in this field (Cunningham & Francis, 2001). Additional developments include Heliz. "Helix an open source digital-media delivery platform designed to let companies build custom applications that stream any media format on any major operating system to any computing device" (RealNetworks, 2003).

In addition to creation, editing, and encoding tools, a number of companies provide the means to check the bandwidth of a video file as it is being encoded and compressed. Terran has a function that will graph the data rate of a video. Macromedia Flash also has a tool called Bandwidth Profiler that will graph streaming data rates (Kennedy, 2000).

DEVELOPMENTS AND TRENDS

There have been some developments and trends occurring in the streaming video field. Most of these are geared to providing new technology, increasing network bandwidth, improving video quality, and competing against television. In the new streaming technology area, there have been some recent efforts to develop and introduce streaming technology into the wireless networks. Toshiba has developed a chip with an MPEG-4 encoder that allows third-generation mobile networks to support two-way videoconferencing (Williams, 2001).

Also, another player in this wireless area is Thin Multimedia, a company that specializes in wireless multimedia streaming and video messaging (Segal, 2002b). They provide software tools for encoding, decoding, authoring, messaging, and streaming that can be installed on second-generation mobile networks. They are also developing tools for third-generation mobile networks. Some of the features of their products include streaming video on a cell phone and having the capability to tap into a live feed from a Web site (Segal, 2002b). Another application that Thin Multimedia has is video mail product. With this product, people can create videos of themselves, while on a Web site and using a webcam, and then send them to someone else's cell phone (Segal, 2002b). The video screens that display on the phones are small, 112×96 pixels. According to Thin Multimedia, certain media forms, such as movies and wide trailers, may not work well. But, other forms such as video mail, traffic reports, and news do work well (Segal, 2002b).

In the area of network bandwidth and performance, the business sector has developed strategies to deal with the network bandwidth and congestion issues. One of the strategies involves the development of Content Delivery Networks (CDNs), which are networks that have the infrastructure and technology to enable a faster and more consistent delivery of streamed media to users. The goal is to reproduce media content and deliver it to the user in an efficient and straightforward manner (Cunningham & Francis, 2001).

There are several commercial CDNs established that offer their services to Internet users. One of the key players is Akamai, with their FreeFlow technology. There are other players that have networks and technology to deliver media to users. Both players claim a tenfold increase in speed by distributing media content to their world-wide networks of servers (Cunningham & Francis, 2001). In addition to the commercial CDNs, some efforts to experiment with this concept are being undertaken by research groups.

The major streaming technologies are continuously working to extend their products to provide digital rights management (another phrase for copy protection) features, interactivity, e-commerce hooks, and better video quality (Bennett, 2002). For example, Microsoft has developed a new streaming format called ActiveMovie Streaming Format (ASF). This format allows multiple data objects to be combined and stored in a single synchronized multimedia stream. The data objects include audio, video, still images, events, URLs, HTML pages, and programs (Bennett, 2002). This format supports digital rights management and pay per view, and, as mentioned before, it is SMIL compliant, which will allow Web authors to create clickable movies (Bennett, 2002). RealOne is working to develop partnerships with other companies, including Apple and Microsoft, to provide greater flexibility in streaming different media formats.

Streaming media companies have turned their attention to the television market. Many of the companies have been developing technology designed to handle programming applications. And some companies have already begun offering broadband services to users that have high-speed Internet access (Tanaka, 2000). For example, MeTV.com and LikeTelevision.com currently offer opportunities to view TV programs, movies, and so forth.

Some companies are forming alliances or partnerships to build streaming video networks designed to provide broadcast-scale streaming media. It seems that the trend for streaming media providers will be to make broadband content delivery available for high-speed users, and retain low bit rates for dial-up users (Tanaka, 2000).

CONCLUSION

In reviewing the above materials, it become clear very quickly that streaming video is a complex, technical process. Besides the sheer complexity of streaming video, there are other issues such as copyright usage.

However, even with technical complexity and inconsistent information, streaming video is an exciting topic. It does provides advantages to user by allowing them to begin playing video without having to completely download it beforehand. And it is surprising to see the number of applications available and the large number of uses for the technology.

In reviewing the history of streaming video, it became obvious how closely streaming video and TV are tied together in their technologies and in new developments in the information and media areas, and how streaming media themselves would have not come into being if it were not for the development of the Internet. Interestingly enough, the data transmission or bandwidth limitations of the Internet remain one of the biggest challenges to streaming video over a network. The Internet simply was not designed for streaming media.

With the bandwidth limitations of the Internet in mind, a lot the streaming media technology is focused on ways to efficiently deliver video over a network. First, there is the raw video that must be captured and digitized into the appropriate input file format. Then the video must be encoded and compressed into the proper streaming format. Next, the video is delivered over the Internet from a special server, called a video server. The user then receives and plays the video. The process sounds simple but the actual functions are very complex, as can be seen from the discussion of compression techniques and algorithms that was included in this chapter.

In facing these technical challenges, the major streaming technologies, RealOne, QuickTime, and Windows Media, have implemented some very good tools that make it possible for a person without streaming media expertise to create, encode, and play simple videos. Also, these technologies have continued to improve and expand their products and have attracted even more users, as well as businesses and educational institutions.

With the continued growth of the streaming video area, other players have entered the market and provided specialized tools for editing and managing bandwidth requirements. Some players offer services for hosting streaming videos, and others provide consulting services for the entire process all the way from creating to playing videos. Many of the players are working to improve the network delivery process, in order to improve the efficiency of streaming video and improve the quality of the video.

GLOSSARY

Algorithm A procedure for solving a mathematical formula.

Bandwidth The amount of information that can be carried through a phone line, cable line, or device.

Buffering Compensating for a difference in rate of flow of data, or time of occurrence of events, when transferring data from one device to another.

Digitizing A process of converting any graphic medium to digital format, so that computerized equipment can read, store, transmit, and recreate it.

Encoding Conversion of input formats into proprietary streaming formats for storage or transmission to a decoder for streaming media.

MPEG Motion Pictures Experts Group, an international organization that developed standards for the encoding of moving images.

PNM Progressive networks media, an older protocol.

RTP Real-time protocol, one of the most commonly used protocols for streaming media on the Internet.

Streaming Video A sequence of "moving images" that are sent in compressed form over the Internet and displayed by a viewer as they arrive. Streaming media are streaming video with sound.

Webcasting Videos of specific live events are shown at a predetermined time to many viewers. Webcasts are

usually shown in real time; typically, there is no prerecorded material.

CROSS REFERENCES

See *Multimedia; Video Compression; Webcasting.*

REFERENCES

Bennett, G. (2002). Internet video trends. Retrieved May 4, 2002, from The Techno Zone Web site: http://thetechnozone.com/videobuyersguide/Net_Video_Trends.htm

Compaq Computer Corporation (1998). Streaming video technology. Retrieved December 31, 2001, from White Papers Web site: http://www.itpapers.com/cgi/psummaryIT.pl?paperid = 4045&scid = 191

Cunningham, D., & Francis, N. (2001). An introduction to streaming video. Retrieved April 20, 2002, from Cultivate Interactive Web site: http://www.cultivate-int.org/issue4/video/

DoIt & WISC (2002). Streaming media [tutorial on streaming media developed by DoIt and University of Wisconsin-Madison]. Retrieved April 20, 2002, from http://streaming.doit.wisc.edu/tutorial/

Filippini, L. (1997). MPEG-1 audio. Retrieved April 20, 2002, from CRS4 Web site: http://www.crs4.it/~luigi/MPEG/mpeg1-a.html

Fischer, B., & Schroeder, U. (1999). Part 3—Video formats and compression methods. Retrieved September 24, 1999, from Tom's Hardware Guide Web site: http://www6.tomshardware.com/video/99q3/990924/video-3–02.html

Fortner, B. (2002). Section 1: Points from the past; History of television technology. Retrieved March 25, 2002, from Communication Using Media Instructor's Notes: http://www.rcc.ryerson.ca/schools/rta/brd038/clasmat/class1/tvhist.htm

Graham, J. (2002, March 5). Video on demand's supply grows. *USA Today*, Section D, p. 6.

Inventors Online Museum (2002). Inventing television. Retrieved March 18, 2002, from Inventors Online Museum Presents History of the Invention and Inventors of Television: http://www.inventorsmuseum.com/television.htm

Kennedy, T. (2000). Don't be scared of bandwidth math. Retrieved January 12, 2002, from Streaming Media World Web Site: http://www.streamingmediaworld.com/symm/tutor/bandmath/index.html

Kennedy, T. (2001). Streaming basics: Editing video for streaming. Retrieved March 31, 2002, from Streaming Media World Web site: http://smw.internet.com/video/tutor/streambasics2/

Leiner, B. M., Cerf, V. G., Clark, D. D., Kahn, R. E., Kleinrock, L., Lynch, D. C., Postel, J., Roberts, L. G., & Wolff, S. (2000). A brief history of the Internet. Retrieved January 1, 2002, from the ISOC (Internet Society) Web site: http://www.isoc.org/internet-history/brief.html

Microsoft.com (2000). Key concepts in Windows media technologies. Retrieved December 31, 2001, from http://msdn.microsoft.com/library/en-us/dnwmt/html/introwmt7–2.asp?frame = true

Microsoft.com (2001a). A brief history of the Internet. Retrieved January 1, 2002, from Microsoft Insider,: http://www.microsoft.com/insider/internet/articles/history.htm

Microsoft.com (2001b). The digital media revolution. Retrieved December 31, 2001, from Windows Media Technologies: http://www.microsoft.com/windows/windowsmedia/overview/default.asp

RealNetworks (2000). Chapter 6: What is streaming media and how does it work? Retrieved January 12, 2002, from Real Networks Web site: http://service.real.com/help/player/plus_manual.8/htmfiles/whatisrpp.htm

RealNetworks (2002). Announcing Helix on July 22, 2002. Retrived April 16, 2002, from http://www.realnetworks.com/solutions/leadership/helix.html

Segal, N. (2002a). Sorenson Media: Video compression software. Retrieved May 4, 2002, from Streaming Media World Web site: http://streamingmediaworld.com/video/docs/sorenson/

Segal, N. (2002b). Thin multimedia: Wireless streaming video. Retrieved May 4, 2002, from Streaming Media World Web site: http://streamingmediaworld.com/videos/docs/thin/index.html

Strom, J. (2001). Streaming video: A look behind the scenes. Retrieved May 4, 2002, from Cultivate Interactive Web site: http://www.cultivate-int.org/issue4/scenes/

Tanaka, K. (2000). Motion pictures on the Net: Streaming media industry, technology, and early adopters. Retrieved January 1, 2002, from Internet Society Web site: http://www.isoc.org/inet2000/cdproceedings/4c/4c_2.htm

Vantum Corporation (2001). In-depth comparison of video CODECs [White Paper]. Retrieved March 28, 2002, from Vantam Corporation Web site: http://www.vantum.com/pdf/codecs.pdf

Videomaker Magazine (2001). Streaming video primer. Retrieved December 31, 2001, from Chaminade College Preparatory Web site: http://www.chaminade.org/mis/Articles/StreamingVideo.htm

Williams, M. (2001). New chip could bring video to mobile phones. Retrieved January 17, 2001, from CNN.com Web site: http://www.cnn.com/2001/TECH/computing/01/17/mpeg4.chip.idg/index.html

Virtual Enterprises

J. Cecil, *New Mexico State University*

INTRODUCTION

> A flexible manufacturing facility in a remote location is in full operation. Machines are busy producing parts, automated robots are helping to assemble them into finished products, and conveyor systems are transporting them to a packing and shipping bay. Above the din of the metal cutting operations and assembly activities, a human manager can be seen in the background. Required design information is routinely collected from distributed customers in different parts of the globe and used by process planning know-how software (migrating from another computer located in a neighboring country) to help identify the manufacturing steps involved, then detailed machine instructions are generated by software from yet another partner's computer located in a neighboring state. From time to time, a resident control computer flashes a signal informing the human manager of the parts being assembled and other work in progress including those being packed and shipped.

This scenario provides a snapshot of the future of manufacturing in this country and throughout the world. With remarkable advances in information technology, computer networks (especially the Internet), and manufacturing integration, the achievement of a truly global manufacturing enterprise seems to be within reach. Smaller and mid-sized enterprise in remote parts of the world will increasingly become part of this revolution by forming partnerships with larger organizations. The notion of such "virtual" partnerships, in which distributed organizations form "virtual teams" and develop products for a changing customer-driven market forms the basis of virtual enterprises (VE).

This chapter provides an overview of the concepts, techniques, technologies, and issues that need to be understood, adopted, and studied in the quest to realize a truly virtual enterprise-oriented approach to product development. The underlying theme is the role of the Internet in the design and realization of such virtual enterprises. Some of the challenges and hurdles, which may be encountered by industrial organizations during implementation, are also delineated in this chapter. Other sections of the chapter include discussions of some of the Internet-based, computer-based frameworks for VE realization, modeling, and communication techniques for VE collaboration and a summary of various industry and university projects related to this subject area.

Concept of a Virtual Enterprise

Today, the concept of a virtual enterprise is being widely heralded as a collaborative partnership for the future as it holds distinct advantages and benefits for organizations worldwide. Formally, "a VE can be described as a consortium of industrial organizations, which come together to form temporary partnerships to respond quickly to changing customer demand" (NIIIP, 2002). In a VE, the partner organizations are geographically distributed, possess diverse skills and resources, and collaborate virtually to produce a final product (Figure 1). In a traditional (non-VE) enterprise, the team members are co-located physically in one specific site and usually belong to the same organization.

In this decade and beyond, it is predicted that growing product complexity and resultant diverse skill requirements underscore the need for organizations to work together as a VE. More importantly, such a collaborative framework will enable the harnessing of remote and far-flung manufacturing facilities (and resources) and create new opportunities for these remotely located organizations who can becomes partners and pillars of the American and international industrial base. Small and medium-sized "mom and pop" operations with specialized capabilities can link with the industrial mega-giants or with other similar-sized enterprises to produce a diverse mix of products beginning to typify the evolving global market. For this reason and several others, American and other international industrial organizations have shown keen interest in virtual enterprise-related principles and practices.

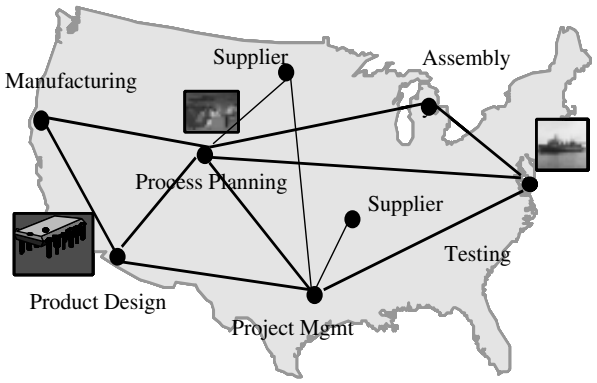


Figure 1: Intra-nation VE collaboration.

Characteristics of a Virtual Enterprise

The concept of a VE is not new. However, very few partnerships can truly claim to have functioned as a virtual enterprise. In this context, it is important to highlight a few aspects related to the historical development of VEs. There are various definitions and connotations of a "virtual enterprise." Numerous reports and articles in newspapers and magazines tout the "functioning" and "success of long running implementations" of VE-based practices. A casual reader who peruses the numerous VE-related articles (that have appeared in the media) might conclude that VE implementation is commonplace and such a model has existed for decades. The key misconception relates to the degree of "seamlessness" of the information exchange in a VE. The questions that must be asked by every reader interested in this topic are as follows:

(1) Is the information merely being exchanged by two partners or organizations by e-mail or the World Wide Web?
(2) Is the information being re-processed or used to perform some engineering or technical task?
(3) Is the information exchange seamless without manual re-keying of information?
(4) Is the information sharing mainly restricted to keeping business partners informed of new or occurring activities?

There are distinctive differences between partnerships that function as "true VEs" and others that are "quasi-VE." The criteria (which can serve as a litmus test) by which organizations can claim that they have implemented VE-based practices include the following:

The partners involved must belong to different organizations and have different core areas of expertise (for example, Company P may have manufacturing expertise while Company F may possess skills in testing products);

The partners must be geographically distributed;

The computer platforms used by the partners must be heterogeneous in nature (having different operating systems such as Windows, UNIX, or Macintosh);

The software systems used in the collaboration must be heterogeneous and be implemented in different programming languages (such as C and Java); and

The information exchange must be through electronic means and must be seamless.

A quasi-VE is a partnership where one or many of the above mentioned criteria holds true (but not all). Most "global" enterprises today function in a manner that can be referred to as 'quasi-VEs'. The most difficult criterion to adhere is the ability to exchange information seamlessly. "Seamless" (in this VE context) refers to the automated translation of information from one data format to another without any manual intervention or re-keying of information. Many market-leading organizations exchange design and engineering information by e-mail and then spend substantial time attempting to extract information by manual means or by using a variety of software in a sequence of tasks. The major drawback of such approaches is that they negate substantially the advantages these organizations intended to accrue by adopting a VE-based model of collaboration. Valuable time and resources are lost by the inability to exchange information from one data format to another. This is perhaps one of the greatest challenges facing organizations interested in adopting a VE model. While the casual reader may conclude that organizations communicating through the Internet in some collaborative manner are functioning as a VE, it is important to underscore the fact that this is a popular misconception. Simply using electronic means (such as the Internet) to exchange information does not guarantee greater productivity. Possessing the ability to use the information (obtained from a partner) immediately in an accessible format to perform a specific target activity is the key to realizing the benefits associated with a VE-based approach. This ability contributes toward the partners being "agile" in a highly competitive environment, where customer needs keep constantly changing. When these needs change, a VE partner may decide to team with a different group of organizations, who may be using a different set of computer tools and data formats. After the virtual teams and responsibilities have been identified, a major issue that must be addressed is the nature of the information and data being exchanged. If this issue is not adequately addressed, then the distributed partners will lose substantial time in accomplishing their collaborative activities quickly. Consequently, the role of computer architectures that support the seamless exchange of information and the importance of data exchange standards needs to be better understood.

Quasi-VEs began appearing in the late 1980s and early 1990s. These include companies such as McDonnell Douglas Aerospace and Ford, among others. Today, many companies (including Boeing) are beginning the migration toward being a true virtual enterprise. Most of them are facing challenges in exchanging information seamlessly (which is a major technical problem) as well as in developing mutual trust with new partners (which is a cultural problem). In the future, successful VEs would be those organizations that emphasize the use of structured proven methods to aid in tasks such as virtual

Table 1 Typical Characteristics of VE Partners in a Satellite Development Domain

Company name	Location	Skills	Responsibilities
Design Corporation	Houston, TX	Space system designers	Design the propulsion system
ASSEMTEC	San Diego, CA	Fabrication specialists	Build various satellite subsystems
Process Consultants, Inc.	Tampa, FL	Manufacturing engineering	Develop plans to manufacture and assemble satellite and its parts
Satellite Design, Inc.	Houston, TX	Space system designers	Design the cold gas, command, and other modules
Agile Integrators	Colorado Springs, CO	Integration specialists	Integration with launch bus and testing
Project Management Consultants	Dayton, OH	Project management	Manage entire product development, budget, oversee task progress, etc.

team formation and are willing to adopt leading-edge technologies that facilitate collaboration among distributed teams. The trend to adopt common information exchange standards such as XML (extensible markup language) and STEP (standard for the exchange of product design specifications) will continue. Today, most manufacturing organizations have indicated their disappointment at emerging standards of data exchange. While there are many technical challenges in adopting data exchange standards, the major resistance to this adoption comes from an unwillingness to change. The emergence of open architecture-oriented practices and the success of organizations today that have embraced such standards will have a definite impact on future trends and practices.

Types of VEs

There are broadly two major categories of virtual enterprises, inter-nation and intra-nation VEs. Inter-nation VEs (or simply international VEs) are those whose members extend beyond national boundaries. For example, consider the electronics-manufacturing domain. Project integrators and design partner organizations may be located in California while process engineering team members and resources are in Texas; in addition, the actual assembly and manufacturing activities can occur in various countries in Asia (such as Taiwan or Singapore). In an intra-nation VE, a consortium's partners are within a specific nation's boundaries. Figure 1 illustrates the concept of an intra-national VE. Another example of VE partnerships and skills is provided in Table 1. In both categories, the Internet can serve as the communication backbone linking the VE members. There needs to be a demarcation between companies who merely have subcontractors in various parts of the world (who may manufacture or assemble parts of a final product) and companies who use the Internet to exchange and share information that directly influences the collaborative activities involved. A manufacturing giant based in California may claim to be part of a global network and yet not function as a true VE. Globalization does not mean just using the Internet or any other electronic means to exchange information. Numerous organizations claim being part of a global network and function more as quasi-VEs and in some situations, mainly subcontract a portion of their activities because of lower manufacturing and other costs. Supplier chain management is one domain in which adoption of Internet-based approaches has proven to be successful. Data exchange has been less of a problem in this domain and this has enabled the adoption of Internet-based practices to support activities related to this domain.

Importance of the Emerging VE Model and the Role of the Internet

At the onset of this new millennium, manufacturing organizations worldwide are collaborating and functioning as a virtual enterprise. With revolutionary advances in information technology (IT) and electronic communications serving as catalysts, the Internet has emerged as a powerful integration vehicle for the realization of the global marketplace. Private and government organizations have recognized the potential of the Internet as a VE facilitator and have begun to implement distributed collaboration approaches using the Internet as a backbone. In this context, there have been several research and industry initiatives that have sought to focus on the development of innovative integration frameworks to support distributed collaborative activities. The term "distributed design, planning, and manufacturing" refers broadly to a subset of VE activities, where physically distributed design, planning, and manufacturing resources interact with each other across heterogeneous computer systems (or platforms) to accomplish identified manufacturing tasks. These resources can include personnel (such as design or manufacturing engineers), software tools (used to create design, manufacturing plans, etc), computer systems (on which the software tools reside), and machines (including robots, assembly and metal cutting equipment, etc).

The development of a product can occur in multiple phases referred to as the product development life cycle. Typically, a life cycle (LC) of a product includes conceptualization of a design idea, detailed design and engineering analysis, project planning, manufacturing and assembly planning, supply chain management, manufacturing (or fabrication), testing, service, delivery and, retirement or recycling (Figure 2). In today's global economy, the complex life-cycle activities involved in developing a product are being performed in a distributed manner (see Figure 1). The project teams, software tools, analysis models, and manufacturing resources involved in this cycle are also becoming increasingly distributed and are implemented on heterogeneous computing systems, which

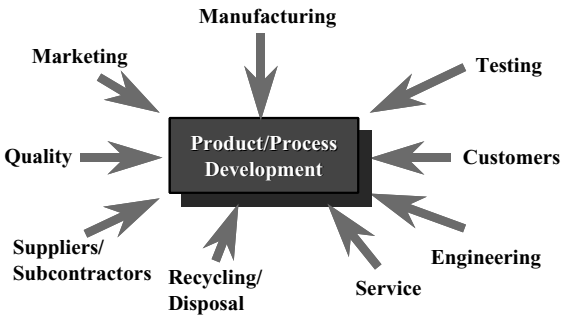


Figure 2: Typical activities in a product development life cycle.

further compounds the existing complex life-cycle integration problem (Cecil, 2001).

Potential Benefits of Adopting the VE Model

The major benefits of adopting a VE-oriented model include (a) the ability to respond rapidly to a newly identified customer need for a product and (b) cost savings. A VE framework will allow Company A to partner with a design specialist (Company B) and a few other organizations to quickly come up with a way to manufacture a product X. After a product demand diminishes or disappears, these VE partners can disband. When another market need (such as a new product) is identified, Company A can form another VE partnership with yet another diverse group of organizations and become involved in developing a new product Y. By becoming temporary partners for a product's life cycle, Company A becomes more agile in being able to serve customers in today's changing global environment. A virtual enterprise-oriented approach will provide organizations the capability to be agile in today's customer-driven global market. The term "agility" is meant to indicate the ability to keep up with customers' changing needs. For example, a cheetah is an agile animal. When its prey changes direction, the cheetah's agile nature enables it to change direction and still continue the chase in pursuit. The Internet along with other technologies is enabling numerous industrial organizations to become more agile. When companies adopt a VE-oriented approach coupled with the power and scalability of the Internet, they will be able to form national and international partnerships to produce low-cost, high-quality products.

CREATION OF VIRTUAL ENTERPRISES

Problems and Challenges in the Creation of VEs

The challenges facing VE implementation can be grouped under technical and cultural. One of the major technical problems is achieving seamless exchange of information as well as the integration of the myriad of activities involved in designing and building products. With the ever-increasing use of the Internet by industrial organizations today to exchange technical and business information, more advanced and sophisticated IT-based frameworks and approaches that will support accomplishment of complex life-cycle activities seamlessly are under development. To improve the productivity of VEs, a variety of issues, especially those dealing with the role of the Internet as a VE facilitator, need to be addressed. Internet-based computer frameworks and architectures must be carefully evaluated with respect to their ability to support realization of VE goals and objectives. Some of the criteria can include the following: Using computer framework A, can VE partners quickly respond to customer demands? Does the proposed Internet approach facilitate seamless exchange of information including engineering, planning, and other life-cycle data? Can information and software systems communicate across heterogeneous systems (such as Windows, Macintosh, and UNIX environments)? At any given time, are the VE partners aware of each other's task accomplishments, their work-in-process (WIP), and the rate of progress toward achieving their overall target production?

The cultural problems relate to the ability to trust new partners, adopt new ways of collaboration, and interact effectively with team members who have less face-to-face interaction during collaboration.

Technologies and Frameworks for the Realization of VEs

The Internet, by far, is the most versatile communication vehicle that can be used to create and manage VEs. It is being widely used by business enterprises globally to exchange information in all phases of a product's life cycle. The Internet can be viewed as "a network of networks" that is scalable and can connect remote corners of our world. The two most widely used protocols of the Internet are the transfer control protocol (TCP)/Internet protocol (IP) (commonly referred to as TCP/IP) and the hypertext transfer protocol (HTTP) (which is better known as the World Wide Web). The TCP/IP was developed nearly 30 years ago and it's the backbone for most of the computer-based communications today. Communication via e-mail, controlled discussion groups on specific topics, and Internet video-based conversations has become commonplace and is replacing the more traditional and expensive telephone discussions and satellite based video conferencing as well. Architectures such as CORBA (discussed later in this chapter) have been developed on protocols such as TCP/IP. Other developments such as the advent of "Internet2" (which will provide substantially more bandwidth and more effective transmission of video graphics and virtual reality-based images) will continue to emerge and mature in response to industrial and educational needs.

Another technology (which has been used widely by a large number of industry giants including Wal-Mart and Ford) is "Electronic Data Interchange" (EDI). Partners involved in an EDI transaction can exchange information from one computer to another directly in a secure automated manner (Cecil, 1996). With the help of translators and transmission standards (national and international), business and technical information including purchase orders, invoices, quotes, and design documentation can be exchanged. Also, electronic funds can be exchanged from one computer to another.

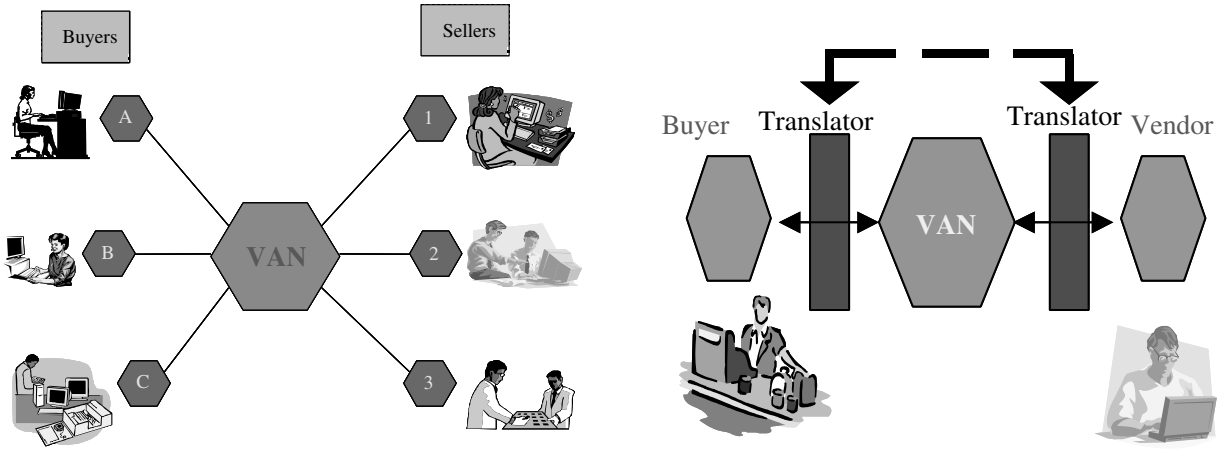


Figure 3: Buying and selling using EDI.

Just as Internet services today are provided by independent Internet service providers (ISPs), EDI services can be obtained through value-added networks (or VANs) (Figure 3). Companies such as GE, IBM, and AT&T provide these EDI services, depending on individual and group needs. EDI frameworks allow for the establishment of electronic bulletin boards where various needs can be posted and subscribers can respond to the business opportunities using EDI. The U.S. Defense Logistics Agency (DLA) is managing the federal government's implementation of EDI. In recent years, most of the VANs have offered a hybrid EDI/Internet-based service where the World Wide Web has been used to exchange EDI-based documents between customers and businesses. Additional information on EDI can be obtained from a number of sources, including books and the Internet (ECRC, 2002).

Computer Architectures and Technologies That Support the Realization of VEs

When teams and resources are distributed and linked via the Internet, as mentioned earlier, the major problem is the ability to communicate across heterogeneous computer platforms. The various software modules and systems used to accomplish engineering and business functions can be implemented on various software paradigms or frameworks. One of the more important (yet basic) concepts in the realm of software computing is the notion of an "object." An object-oriented software program (or an "object") can be viewed as a discrete software entity that contains some "data" that can be manipulated using certain functions or operations. In general, such an object has several advantages over traditional software entities built using non-object-oriented languages such as Fortran. Object orientation provides three distinct advantages including ease of maintenance, ease of change, and less time to create. Objects can be reused, and in most cases, they provide a basis that can be extended. Objects can be created from a template. The template used to create a group of objects is termed a "class." In a manufacturing or any other enterprise, most objects model real-world entities. Software entities can send messages to objects with specific requests; the objects, in turn, send their responses through messages.

In an Internet-based VE, the various software entities that are distributed can be viewed as engines and turbines working together to propel a given enterprise. Using distributed object-computing methods, communication among the physically distributed software systems is possible. Distributed computing allows objects to be distributed in an heterogeneous manner across the Internet by extending object-oriented programming concepts so that these distributed objects behave as a unified whole. These objects can reside in their own address space outside of an application and be distributed on different computer platforms linked via the Internet; however, they will behave as if they were local objects. There are several ways to implement a distributed computing environment, which is a key requisite to realize a fully functional VE. Three of the most popular paradigms and approaches are discussed in the following sections: the common object request broker architecture (CORBA) from the Object Management Group (OMG), the distributed component object model (DCOM) from Microsoft, and Jini technology from Sun Microsystems. Among these three, CORBA and DCOM are architectures and can be compared. Jini is built on top of the Java language and has become popular as it enables systems to function as a federation of services.

The Common Object Request Broker Architecture (CORBA)

The Object Management Group is the world's largest computer industry consortium whose mission is to define a set of interfaces for software to be interoperable. OMG is a nonprofit organization with around 750 members. The OMG provides a structure and a process through which its members can specify technology and then produce commercial software that meets those specifications. CORBA is an industry consensus standard that defines a higher-level facility for distributed computing. The distributed environment is specified using an object-oriented approach, which masks the differences relating to object location, type of operating system or computing platform,

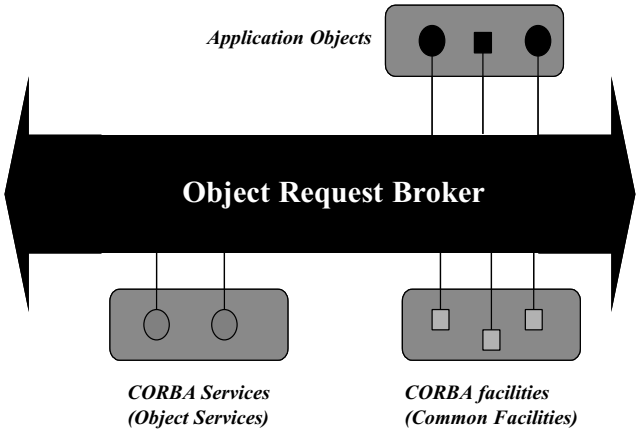


Figure 4: The object management architecture.

and programming languages (used for implementing objects). CORBA supports interoperability by using a well-defined set of interface specifications (Mowbray & Zahavi, 1995).

CORBA is a specification for an application level communication infrastructure. It can be viewed as a peer-to-peer distributed computing facility capable of supporting seamless communication and information exchange, which is the cornerstone of VE implementation. In a CORBA context, all applications are viewed as objects. These objects are capable of assuming dual roles as a client and a server. A client invokes a service to which another object (referred to as the server) responds. In general, CORBA allows more flexibility from a computer architecture point of view when compared to pure client–server frameworks allowed by remote procedure calls (RPCs).

While CORBA connects objects, the realization of a VE requires any supporting computer architecture to support enterprise integration. The object management architecture (OMA) seeks to address this VE requirement (Figure 4) and is based on CORBA. The major components of OMA include the object request broker, CORBAfacilities, CORBAservices, and application objects. The object request broker (ORB) can be viewed as a communication infrastructure capable of relaying object requests and responses transparently across distributed computing environments such as the Internet. The CORBAservices (also referred to as object services) provide lower level functionality such as life-cycle services (such as object creation and notification) and include providing access to online transaction processing (OLTP, the widely used application for business accounting) and a "yellow pages" type of trader service (where objects can offer their services at specified costs).

The CORBA facilities (referred to as common facilities) provide services for applications and have two major segments, horizontal and vertical. The horizontal facilities can be widely used in a variety of industrial domains and markets and include user interface, task management, information management, and systems management. The vertical facilities deal with standardizing management of information pertaining to specific industrial domains such as healthcare, manufacturing, and financial. An example of the horizontal facility is the compound document management facility, which allows applications a standard way to access the various parts of a document. Using this, a vendor can build tools for manipulation of a part of the document (for instance, a three-dimensional image), which can then be marketed without having to develop the functionality from scratch.

The OMA and CORBA both facilitate the creation and management of Internet-based software tools used in a typical VE-oriented product development cycle. They allow engineers and managers in design, manufacturing, and other areas to take advantage of a wide variety of software tools implemented in various programming languages and on diverse operating systems (from a Macintosh to a UNIX system). IT specialists and software programmers especially will benefit from using OMA/CORBA because they provide a sophisticated mix of easy component accessibility and transparent distribution; further, code can be reused in new applications or can be modified incrementally to suit the increasing scope of applications in a VE. Tools and modules can be developed in a variety of languages. For example, a Java-based program can be used to perform a manufacturing optimization task (in your VE) while a C++ module can be built to create a user interface that is accessible to everyone in a VE.

The key to object interoperability in CORBA depends on the definition of contracts termed "interfaces." Each object has an interface (which is public) and an implementation (which is private). The services of each object are expressed as a type of contract in this interface, which provides two very important functions: (a) it informs other client objects in the VE of the services it provides as well as specifies how to be "called" (or invoked); and (b) it allows the IT infrastructure to understand the specific manner in which it will send and receive messages. The latter allows for data translation between client and server objects. Each object in the VE also requires a unique identifier or handle, which can be used to direct messages. The interfaces of all the objects can be expressed in a neutral language called the interface definition language (IDL). In CORBA, the IDL definition for all objects is stored in an interface repository. This repository is useful to achieve object interoperability, which is a key issue in the creation of systems supporting the functions in an Internet-based VE.

In CORBA, IDL is the means by which a particular object implementation tells its potential clients what operations are available and how they should be invoked.When an application object is created in a specific language such as C or C++ (or any other language that can be mapped to IDL), each object's interface can be defined in IDL. There are two ways for distributed objects to be linked in CORBA: using static IDL interfaces or using the dynamic invocation interface (DII). When using static interfaces (during implementation), the IDL specifications are compiled for each interface into header, skeletal, and stub programs for linking the Internet-based distributed applications. When a client object (such as a task manager located in VE site 1) invokes an operation on an object reference, it links using *stubs* (which are generated automatically from an IDL compiler for the language and

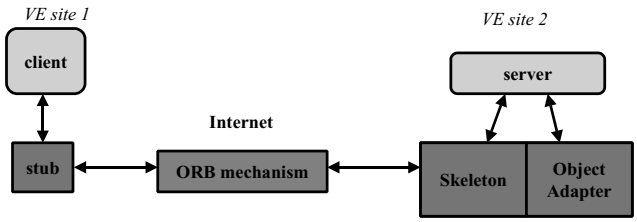


Figure 5: Linking in a VE using stubs and skeletons.

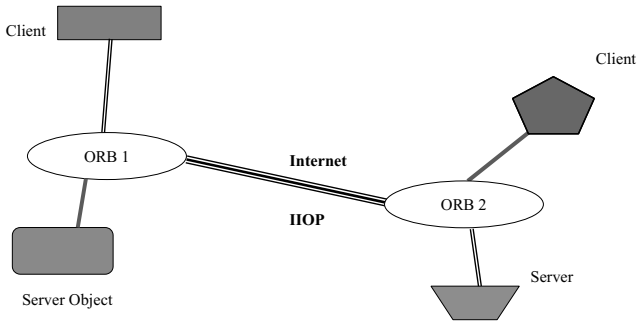


Figure 6: Clients and server objects communicating via the Internet using IIOP.

ORB environment the client is implemented in). These stubs convey this information to the target object via the ORB (Figure 5). From the client's perspective, the stub acts like a local function call. After the stub interfaces with the ORB, the ORB encodes and decodes the operation's parameters into formats suitable for transmission (this is termed "marshaling"). On the server object side (the server may be a manufacturing analysis object or any other application object located in another VE site 2), the information is "unmarshaled" and a skeletal program interfaces with the ORB through an object adapter. This skeleton can map the request back to the implementation language of the server object (the skeletal program is obtained using an IDL compiler the server is running on). When the ORB receives a request, the skeleton in essence performs a call-back to the server object. After the server completes processing the request, the results are returned to the client via the skeleton/stub route (along with any reported errors or problems encountered).

The static interface approach does not support the dynamic use of newly created objects once they are introduced into the VE's network. The dynamic invocation interface provides this function and enables a client to discover new objects and their interfaces, retrieve their interface definitions, construct and send requests, and receive the associated response from objects on the Internet.

In CORBA, the encapsulation properties of the various software objects enable location transparency in the VE. In an encapsulated component, there are two parts: the *public* interface (presented to the outside world) and the *private* implementation (which is appropriately hidden from view). When a client sends a message, the invocation is sent to the ORB (and not the target server object); the ORB routes the message to the destination or server object. Consequently, the location of the object within the virtual enterprise does not matter. How the results were obtained is of concern only to the server component (or object) and the client need not know how the server processed its request. The interfaces can be viewed as contracts by an object. By using IDL, which allows the interfaces to be specified in a neutral language, it is possible to separate interfaces from the implementation. The mapping from IDL to languages such as C, C++, Java, and Smalltalk (among others) can be achieved by enabling various resources in a VE to be implemented in a variety of programming languages running on heterogeneous platforms ranging from UNIX to Windows. By using IDL to specify interfaces, different VE partners and team members can independently implement different parts of a distributed system, which can be used to accomplish target life-cycle activities (ranging from design through testing of a product). When application objects in a VE are moved from one site to another in a VE, the ORBs can use the object references to locate them. In general, the client does not know whether a target server object is local or distributed (and linked using the Internet).

In any typical VE-oriented implementation, each Web site of a company partner will have an ORB. ORBs (whether implemented in the same language or in different languages at each site) can communicate to each other using the Internet Inter-ORB protocol (IIOP) (Figure 6). The IIOP is the general inter-ORB protocol (GIOP) over the TCP/IP, which is mandatory for CORBA 2.0 compliance. The IIOP is based on the TCP/IP, which is the most popular transport mechanism available today and is the protocol of the Internet. Interoperable object references enable invocations to pass from one (language) ORB to another.

Security is an important issue in any distributed approach; use of various computers introduces issues of consistency and trust between them. In a distributed system, information is in transit and is more vulnerable to outside "attacks." CORBAsecurity is part of the CORBAservices and is flexible and can be modified to suit different security needs (Seigel, 2000). The building blocks of CORBAsecurity include identification and authentication of principals, authorization and access control, delegation, non-repudiation, cryptography and maintaining availability, and performing security auditing to maintain user accountability.

Personal computers (PCs) constitute the major segment of desktop computer systems used in industry today. Using OMA/CORBA, PCs can become powerful participants in the design and functioning of VEs linked using the Internet. Each ORB can be from a different vendor and implemented in diverse programming languages. ORB products can differ greatly yet conform to the OMG specifications and guarantee interoperability over the Internet. Some of the ORB products include ObjectBroker (from the formerly known Digital Equipment Corporation), the SOM product set (from IBM), the DAIS product set (from ICL), Orbix (from Iona Technologies), Distributed Smalltalk and ORB Plus (from Hewlett–Packard), and the NEO product family (from SunSoft, Inc.). Additional information on CORBA and its specifications can be obtained from the OMG Web site (OMG, 2002).

The Distributed Component Object Model (DCOM)

The Microsoft distributed component object model (DCOM) also referred to as "COM on the wire" uses a protocol called object remote procedure call (ORPC) (Microsoft, 2002). DCOM is used substantially on the Windows platform. Microsoft provides common object model (COM) implementations for Windows and Solaris platforms while other companies provide implementations for UNIX, Linux, and other mainframe platforms. The Active Group, which is a consortium of vendors interested in the evolution of COM and DCOM, manages both these specifications. COM required that the provider and user of an interface both reside on the same computer. For instance, Microsoft Visual Basic could activate and use Microsoft Excel on the same computer but was not capable of controlling Excel on another computer located on the same local area network or on the Internet. DCOM extends the original COM to support communication among distributed objects on different computers in the Internet (or local area or wide area networks). An interface client can make a request for another interface, which can be provided by an instance of another object, which is on another computer on the Internet. COM's distribution mechanism can connect the client to the server so that the method called from the client is received by the server (or provider) on another computer where it is executed and the return values are returned to the client (or consumer). The distribution is transparent to both clients and servers.

One of the major mechanisms in COM is the activation mechanism, which establishes connections to components and creates new instances of components. In COM, object classes possess globally unique identifiers (GUIDs). Class IDs are GUIDs used to refer to specific classes of objects. In DCOM, the object creations in the COM libraries are enhanced to allow creation on other computers linked via the Internet. To create a remote or distributed object on the Internet, the COM libraries must know the network name of that server (and the class identifier CLSID). Based on this information, a service control manager on the client computer links to the service control manager on the server computer and then requests creation of a new object. DCOM provides several ways to allow clients to indicate the remote server names when a new object is created.

DCOM uses the object remote procedure call, which is a layer on top of the distributed computing environment's (DCE) remote procedure call and interacts with COM's run-time services. A DCOM server object is a piece of software code capable of offering objects of a specific category at a certain time (called run-time). Each server supports multiple interfaces, which can each represent different behaviors. A client in DCOM first acquires a "pointer" to one of the server's (or provider's) interfaces and then makes a call to one of its exposed or public methods. Using the interface pointer obtained, the client objects can invoke the public methods of the target server even though it is located on another computer available on the Internet. The servers in DCOM can be written in various programming languages such as Visual Basic, Java, and C++.

COM uses the remote procedure call infrastructure to accomplish marshaling and unmarshaling. For this, the exact method signature including the data types, sizes of any arrays in the parameter list, and types of structure members must be known. This information is provided in an interface definition language, which is built on top of the industry standard IDL (described earlier in the CORBA section). The IDL files are compiled using special compilers (such as the Microsoft IDL compiler MIDL), which generate C language source files, which contain the code for marshaling and unmarshaling for the interfaces described in the IDL file. The client code is termed the "proxy" and the server object is called the "stub." When the proxy/stub for a specific interface is needed in COM, the interface ID (IID) is identified from the system registry. Whereas CORBA supports multiple inheritances at the IDL level, DCOM does not. ObjectBroker (which is Digital Equipment Corporation's implementation of the CORBA specification) can work with Microsoft's object linking and embedding (OLE) functions for data objects stored on non-Microsoft platforms. The interfaces to Microsoft's OLE and dynamic data exchange (DDE) are available in ObjectBroker. ObjectBroker's OLE network portal can intercept OLE calls on the PC and map them to ObjectBroker messages, which can then be routed to the appropriate server. In addition, ObjectBroker also allows interfacing with Microsoft's Visual Basic, which enables graphical applications to be developed quickly and extends the desktop's capabilities to access information residing on computers linked via any network including the Internet.

Java and Jini Technology

Java remote method invocation (RMI) relies on a protocol called the Java remote method protocol. Java relies on object serialization, which allows objects to be transmitted as a stream (Raj, 2002). The major drawback is that both the server and client objects must be written in Java. However, Java RMI can be implemented on a variety of heterogeneous operating systems (from UNIX to Windows) with one restriction: there should be a Java Virtual Machine implementation for that platform. Internet-based VEs that have no legacy software and are not concerned about the introduction of heterogeneous implementations (in C++ and other languages) can implement Java-based systems. Java has many advantages including being object-oriented and offering ease of programming, modularity, and elegance. Jini (from Sun Microsystems) seeks to extend the benefits of object-oriented programming to the Internet (or any network). Jini is built on top of Java, object serialization, and RMI and enables computers to communicate to each other through object interfaces. Jini provides for a more network-centric environment where computers not possessing a disk drive become more commonplace and interact over dynamically changing networks (such as the Internet). Jini technology provides ways to add, remove, or locate computers as well as services. It can help VE partners build and use a distributed system as *a federation of services* to accomplish their target activities. The set of all available services available on the Internet to the VE will compose this federation with no specific service in charge. Jini's infrastructure provides a way for clients and services to locate each other using a lookup service (which is a directory of currently available services).

Agents and Mobile Agents

Agents have also become increasingly popular and are well suited to support the functioning of Internet-based VEs. Agents can be viewed as software entities, which are semi-autonomous, proactive, and adaptive and which have a long life. They can collaborate with each other to work toward common or independent goals (Deshmukh, Krothapalli, Middlekoop, & Smith, 1999; Krothapalli & Deshmukh, 1999; Lange, 1999). Using an Internet-based framework, agents that can help accomplish a number of tasks in a VE can be designed and created. These tasks can be system oriented (for example, monitoring and notification of task completions) or product/process oriented (for example, generation of a plan to assemble three parts). Agents called "mobile agents" hold enormous potential in revolutionizing the way in which VEs function. A *mobile agent* is capable of replication and autonomous movement from one VE site to another and of performing tasks based on information collected from various sites in a VE. These are in contrast to *stationary agents or objects,* which execute only on the system they reside on and when they need to interact with objects on another system, use methods such as remote procedure calling (RPC). Mobile agents are not bound to their host site or system but can travel (or "roam") among the various computing hosts. Such an agent can transport its state and code along with it to another location, where it can resume execution. Several university and industry projects have highlighted the benefit of using mobile agents in distributed environments (Cecil, 2002a; Lange, 1999). The potential advantages of a mobile agent approach include overcoming network latency (they can be dispatched from a remote control mediator to act locally and eliminate latency necessary for real-time response and control), encapsulating protocols, and offering better performance, increased flexibility, and support dynamic response to changing scenarios. The tactical advantage of performance is gained by sending a component across the network to a VE site, where the work gets completed. The computers in the VE need to be connected only long enough to send the mobile components and later to receive it back. This same concept comes in useful especially for monitoring remote activities in manufacturing and other domains (Cecil, 2002b).

Industry and University Initiatives Related to Internet-Based VEs

Many industrial organizations with remote locations and distributed resources and partners have adopted a VE mode of functioning using the Internet as the communication backbone. The VE model has been adopted in a wide variety of industries including aerospace engineering, airline and travel industry, shipbuilding, computer manufacturing, healthcare, IT systems consulting, banking, electronic commerce, and telecommunications. In general, organizations involved in the service and consulting sectors (who do not produce a physical product for customers but rather provide services) can use Internet-based frameworks to exchange information seamlessly in a distributed collaborative manner. Organizations such as Ford, Boeing, the Sabre Group, Lufthansa, Schlumberger, Pratt and Whitney, Cisco Systems, Raytheon, Harvard University, and NASA Goddard Space Flight Center have adopted CORBA-based architectures to be more agile and customer responsive (CORBA, 2002).

Boeing, the world's largest producer of commercial airliners, uses a CORBA-based framework to integrate its design, manufacturing, and resource management activities. The manufacturing and assembly activities at Boeing involve as many as 3,000,000 individual parts for each aircraft produced. Information integration, inventory management, collaborative sharing of design, manufacturing, and work-in-process data are extremely complex and require a robust distributed IT infrastructure. Boeing's Internet-based VE computer architecture (which is based on the OMA/CORBA model) can support more than 45,000 users and 9,000 concurrent users in various regions across the USA (CORBA, 2001).

The NIIIP project is a national initiative that focuses on developing, demonstrating, and transferring (to interested organizations) the technology to enable industrial virtual enterprises. The National Industrial Information Infrastructure Protocols (NIIIP) consortium is a group of industry, university, and government organizations defining (and have defined) the NIIIP protocols as well as demonstrating their use. Some of the consortium members include CAD Framework Initiative, Digital Equipment Corporation, IBM, General Dynamics Electric Boat, Lockheed Martin Aeronautical Systems Company, National Institute of Standards and Technology, STEP Tools, Rensselaer Polytechnic Institute, Texas Instruments, and the University of Florida. Additional information about the NIIIP reference architecture can be obtained from their Web site (NIIIP, 2002). NIIIP aims to establish standards-based software framework protocols as well as develop software and toolkits (as part of its effort to provide a technical foundation) for implementing virtual enterprises. There are four building blocks of the NIIIP reference architecture, and they include communication (using the Internet), use of object technology, knowledge and task management, and common information model specification and exchange.

Several U.S. federal programs have initiated projects to design collaborative virtual environments. Among these are the Distributed Knowledge Environment of the Department of Defense; Intelligent Collaboration and Visualization initiative of the Defense Advanced Research Projects Agency (DARPA); System Integration for Manufacturing Applications; National Advanced Manufacturing test bed of NIST; Rapid Design Exploration and Optimization (RaDEO), and Agile Infrastructure for Manufacturing Systems (AIMS). A distributed Internet-based process planning system called CHOLA has been developed at New Mexico State University (Cecil, 2001; Cecil, 2002a). CAD files and process planning modules are distributed among heterogeneous environments. Dynamic information (such as design data, equipment capability and availability, and tool availability) from remote locations is used by the VE (which includes ITESM in Monterrey, Mexico, and Penn State University in State College, PA) to generate a process plan for various part designs (Cecil, 2002a). Each site has an object request broker, which acts as a communication infrastructure linking the distributed sites via the Internet. This combined research

and curriculum development initiative was funded by the National Science Foundation as part of the emphasis to introduce emerging engineering concepts and practices to engineering students.

Activities and Phases in the Creation of Internet-Based VEs

There are no structured methods or steps to support the creation of VEs today. However, by addressing the major issues involved to establish, sustain, and function as a VE, industrial organizations can develop their own approaches to successfully adopt VE practices. The major activities in any VE development include the following:

Development of an understanding of a given VE's product(s) and customers;

Identification of the potential VE partners and formation of the product development teams;

Development of an information-based enterprise model of the VE's collaborative activities and tasks;

Design and implementation of an Internet-based distributed software system that will link all VE team members (and possibly customers) and be used to accomplish the various VE activities;

Initiation of a pilot initiative in which partners function as a VE using the implemented Internet-based system; and

Based on performance in pilot initiative, identification and adoption of necessary changes to the overall approach and Internet-based architecture and software system.

Among these activities, the most important activities are the creation of an information-oriented enterprise model and the work associated with the design and implementation of an Internet-based system to support the functioning of a VE (Cecil, 2002c). The use of information-oriented enterprise models are becoming increasingly popular and have their roots in the integrated definition (IDEF) work that originated in the ICAM Program at Wright Patterson Air Force Base decades ago. Apart from older methodologies such as the IDEF-0 and IDEF-3 (Cecil, 2002b; Mayer, 1992), more recent development include the enterprise modeling language (EML, proposed by Virtual Enterprise Technologies) (Xavier & Cecil, 2001). EML was designed with the primary goal of enabling companies interested in creating and participating in VEs to conceptualize and model the way in which partner organizations would interact with each other in accomplishing various enterprise activities. Further, it provides a structured way to propose and refine how to accomplish detailed activities encompassing all or some activities in a product development cycle (ranging from development of a product design collaboratively to the generation of project and manufacturing plans to supplier chain management). EML enables VE team members to describe what activities will be performed and how to accomplish them using available resources and constraints. Using decompositions, detailed models at various levels of abstraction can be built. In EML, the top-level

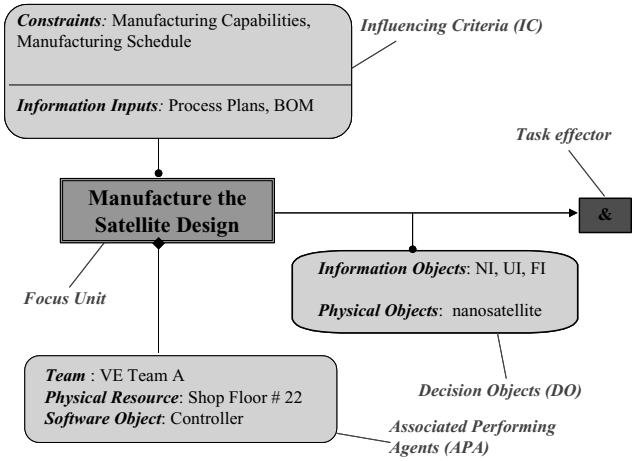


Figure 7: Functional unit representations using EML.

life-cycle activities are captured as functional units. Each functional unit can correspond to an identified life-cycle function such as "create product design X" or "manufacture design X using VE partners." An information-rich description (or model) can be developed using four classes of attributes including influencing criteria, task effectors, decision objects, and associated performing agents. Figure 7 illustrates the main information attributes captured in EML, which include focus units and the four attribute classes. EML provides a structured basis to model VE activities, capture their interrelationships, and specify their accomplishment using temporal precedence criteria.

Influencing criteria help VE team members identify major constraints as well as enable identification of information needed by VE teams or a VE subcontractor to accomplish a certain task. For example, in Figure 7, the information inputs needed to accomplish the target activity "manufacture the satellite design" include process plans and Bill of Materials (BOM). Associated performing agents help model who or which mechanisms will actually help accomplish target activities or subactivities.

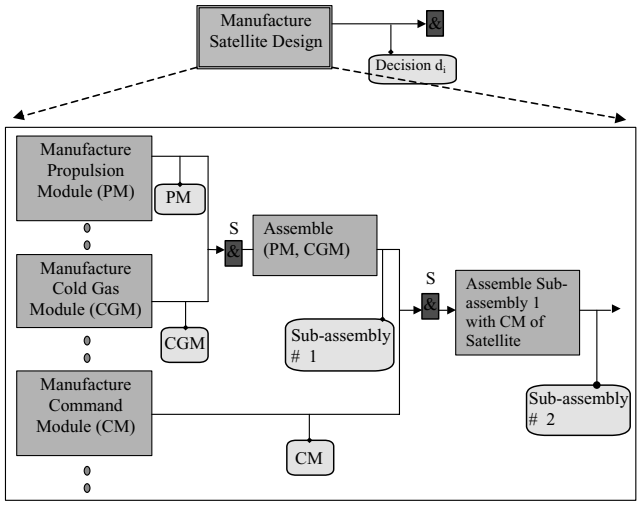


Figure 8: Decompositions of functional activities in EML.

Decision objects represent outputs after tasks are completed. Task effectors enable capturing the temporal logic underlying the task accomplishments (for instance, should Tasks A and B be accomplished concurrently or sequentially?). Figure 8 shows the decomposition of the functional unit described in Figure 7 (the boxes corresponding to influencing criteria and associated performing agents have been omitted for easier reading in Figure 8). In general, enterprise model building of future activities will enable the IT team composed of VE team members to design and build effective Internet-based systems (based on CORBA or any other architecture), which will be the backbone of the various distributed VE activities.

CONCLUSION

The Internet is a powerful vehicle for VEs to be created and deployed. This chapter discussed the major Internet-based approaches, tools, and technologies available today to establish virtual enterprises. Other supporting technologies can also be used on the Internet to promote better communication among distributed team members in a VE. An example of such a supporting technology is virtual reality, which can play a major role in the functioning of VEs (Banerjee, Banerjee, Ye, & Dech, 1999; Brown, 1999; Cecil 2002b, 2002c; Goldin, Venneri, & Noor, 1999). Distributed team members can communicate effectively using this powerful technology from various locations. With the development of Internet2, the use of virtual reality for VE task accomplishments is expected to become more widespread. A key aspect of collaboration, which must be embraced in any VE, relates to the notion of "concurrent engineering" (Mayer, Su, & Cecil, 1997). In such an concurrent approach, the distributed cross-functional teams must consider both product and process design issues simultaneously to ensure reduced costs, shorter development lead time, and higher product quality.

Internet2 is under development by a partnership involving U.S. universities, industry, and government; it is a more advanced network that will link universities, government, and research laboratories for the purposes of collaboration, distance learning, research, health services, and other applications that require high bandwidth between the distributed sites. Internet2 is not intended to replace the Internet; rather, it will complement the capabilities of the Internet by providing additional capabilities such as the development and deployment of advanced network applications and technologies, including substantial increase in the bandwidth. Logistical networking is a new approach for synthesizing networking and storage to create a communicative infrastructure for network multimedia and distributed applications. In May 2002, researchers from Logistical Computing and Internetworking (LoCI) Laboratory at the University of Tennessee, where research in logistical networking research is being pursued, demonstrated "Video IBPster," an application that can deliver video at high performance. The technology used is simpler and less expensive to deploy than current approaches to streaming video. Additional information is available at the Internet2 Web site (Internet2, 2002).

The adoption of the Internet as the vehicle of communication by industries worldwide will continue to grow. As both technology and practices mature, the VE model is expected to become more widespread. There will be more emphasis on the structured design of VE approaches (using information modeling methods), development of effective virtual team formation/interaction methods (Hackman, 1990), and seamless exchange of information. Organizations that adopt a quasi-VE approach will be forced (by increasing competition) to focus on unproductive practices relating to data incompatibility and information exchange. Smaller organizations will be able to form partnerships with larger enterprises and have more access to more market opportunities worldwide.

The Internet has provided a more open approach and increased business opportunities for industrial organizations worldwide. It has transformed the essential manner in which new products are made, created a more customer-oriented environment and radically changed the manner in which people communicate with each other. The Internet is the cornerstone of the information technology revolution. It has created new opportunities and provided groundbreaking avenues to better health, education, and literacy; it has become the de facto communication vehicle of choice by millions worldwide and has allowed us to be closer to each other than ever before. When a group of engineers and programmers began the creation of the Internet several decades ago, they had little idea of its far-reaching impact. Today, thanks to their vision and dedication, people all over the world feel closer to each other, even if oceans separate us. The cliché is true: We are but a cyber-click away from each other. Our world will never be the same again.

GLOSSARY

Computer architecture The functional appearance of a computer to its immediate users or, as in this chapter, the relationship of the various software elements and the manner in which they interact with each other and the processor to accomplish specific tasks. The CORBA and DCOM models discussed in this chapter are examples of two different architectures.

Distributed computing A functional task or activity completed in a collaborative manner by humans and/or software systems that are not co-located but residing on several geographically distributed computers linked via the Internet or any other network.

Flexible manufacturing The ability to manufacture a wide variety of parts using reconfigurable computer-controlled manufacturing equipment.

Mobile agents Software entities that migrate from one computer to another on a network such as the Internet.

Process planning A task or function that identifies the process steps needed to manufacture a given design.

CROSS REFERENCES

See *Client/Server Computing; E-systems for the Support of Manufacturing Operations; Intelligent Agents; Virtual Reality on the Internet: Collaborative Virtual Reality; Virtual Teams.*

REFERENCES

Banerjee, A., Banerjee, P., Ye, N., & Dech, F. (1999). Assembly planning effectiveness using virtual reality. *Presence—Teleoperators & Virtual Environments, 8*(2), 204–217.

Brown, A. S. (1999). Virtual people take on the job of testing complex designs. *Mechanical Engineering, 121*(7), 44–49.

Cecil, J. A. (1996). Electronic commerce and the electronic commerce resource centers. *Marine Safety Council Proceedings—The Coast Guard Journal of Safety at Sea, 53*(4), 10–13.

Cecil, J. A. (2001). *Distributed heterogeneous architecture for manufacturing applications (DHARMA)* (Project Report). Las Cruces, NM: Virtual Enterprise Engineering Laboratory (VEEL), New Mexico State University.

Cecil, J. A. (2002a). *A distributed framework for satellite product development* (Project Report). Las Cruces, NM: Virtual Enterprise Engineering Laboratory (VEEL), New Mexico State University.

Cecil, J. A. (2002b). A functional model of fixture design to aid in the design and development of automated fixture design systems. *Journal of Manufacturing Systems, 21*(1), 58–72.

Cecil, J. A. (2002c). DHARMAM: An university–industry partnership to enhance manufacturing engineering education. Paper presented at *Advanced Manufacturing Institute's International Conference on University & Manufacturing Industry Collaboration,* August 12–13, 2002, Kansas City, Missouri.

Cecil, J. A., Kanchanapiboon, A., Kanda, P., & Muthaiyan, A. (2002, July). A virtual prototyping test bed for electronics assembly. In *Proceedings of the 2002 International Electronics Manufacturing Technology (IEMT) Symposium* (pp. 130–136). Piscataway, NJ: IEEE Service Center.

CORBA (2001). *Manufacturing: Boeing Commercial Airplanes Group*. Retrieved May 24, 2002, from http://www.corba.org/industries/mfg/boeing.html

CORBA (2002). *CORBA success stories*. Retrieved May 24, 2002, from http://www.corba.org/success.htm

Deshmukh, A., Krothapalli, A., Middlekoop, T., & Smith, C. A. (1999, January). *Emergent aerospace designs using negotiating autonomous agents: Laboratory for Fundamental and Applied Research in Multi-agent Systems report*. Amherst, MA: University of Massachusetts.

EDI (2002). Retrieved April 10, 2003, from http://www.edi-information.com/

Goldin, D., Venneri, S., & Noor, A. (1999). Ready for the future? *Mechanical Engineering, 121*(11), 61–70.

Hackman, R. (Ed.). (1990). *Groups that work and "those who don't," creating conditions for effective teamwork*. San Francisco: Jossey–Bass.

Internet2 (2002). Retrieved May 24, 2002, from http://www.internet2.edu

Krothapalli, A., & Deshmukh, A. (1999). Design of negotiation protocols for multi-agent manufacturing systems. *International Journal of Production Research, 37*(7), 1601–1624.

Lange, D., & Mitsuru, O. (1999). Seven good reasons for mobile agents. *Communication of the ACM, 42,* 88–89.

Mayer, R. J. (Ed.). (1992). *Information integration for concurrent engineering (IICE)—IDEF3 process description capture method report* (AL-TR-1992-0057). College Station, TX: Knowledge Based Systems.

Mayer, R. J., Su, C. J., & Cecil, J. A. (1997). Enabling concurrent engineering in the product analysis environment. *Institute of Industrial Engineers (IIE) Transactions, 29,* 791–797.

Microsoft (2002). *Distributed component object model (DCOM)—Downloads, specifications, samples, papers and resources*. Retrieved May 24, 2002, from http://www.microsoft.com/com/tech/dcom.asp

Mowbray, T., & Zahavi, R. (1995). *The essential CORBA, systems integration using distributed objects*. New York: Wiley.

NIIIP (2002). *Welecome to: National Industrial Information Infrastructure Protocols and related projects*. Retrieved May 24, 2002, from http://www.niiip.org

Object Management Group (OMG). (2002). Retrieved May 24, 2002, from http://www.omg.org

Raj, G. S. (2002). *A component engineering cornucopia*. Retrieved May 24, 2002, from http://gsraj.tripod.com/

Seigel, J. (2000). *The CORBA 3 fundamentals and programming object management group*. New York: Wiley.

Xavier, B., & Cecil, J. (2001, August). *Design of an enterprise modeling language* (VETI EML Report). Las Cruces, NM: Virtual Enterprise Technologies, Inc. (VETI).

Virtual Private Networks: Internet Protocol (IP) Based

David E. McDysan, *WorldCom*

INTRODUCTION TO IP-BASED VIRTUAL PRIVATE NETWORKS

Applications of IP Virtual Private Networks

The public Internet plays an important role in many enterprises (McDysan, 2000). Users can exchange information with individuals anywhere in the world via e-mail, Web sites, transaction systems, file sharing, and file transfer. Furthermore, the Internet is a rapidly growing means of conducting business for commercial enterprises. It also provides a means for companies to advertise their goods and services. The Internet can help reduce administrative costs by placing the data entry, verification, and think-time aspects of order entry and service parameter selection in the hands of the end user. This replaces the older, less-efficient paradigm of people in enterprises interacting over the postal system and/or the telephone and facsimile to place orders, update records, and complete business transactions. The Web provides the automated means for the end user to peruse the choices at his or her own speed, requiring the expenditure of energy and time of only one person. Furthermore, careful design of the Web site by experts allows many more people access to the best set of information. In the classic telephone or facsimile method, the level of expertise depended on the particular agent the caller reached.

The tremendous volume of such information on public Web sites continues to grow and increase in quality, based upon real-world experience and user feedback. When the Web site contains enterprise-specific information that, for one reason or another, is sensitive, we call the application an intranet. One level of security is that of user identifications (IDs) and passwords. This is the same level of security used on many public domain Web sites. The next level of security is that of encryption and firewalls, topics covered in the next section. A more challenging activity is the use, by multiple enterprises, of the Internet in a virtual private fashion in an application called an extranet. The premier example to date is probably that of the Automotive Network eXchange, which connects major automotive manufacturers and their suppliers, as described at the end of this article.

In addition to control over who may communicate with whom, as described above, virtual private networks (VPNs) have a number of additional important requirements. Of course, providing verifiable authentication that specific sites and users are part of a specific intranet or extranet VPN is an important requirement. Also, keeping the administrative cost of VPNs under control requires automation of membership discovery in conjunction with this authentication. Furthermore, customer networks will make use of private IP addresses or nonunique IP address (e.g., unregistered addresses). This implies that there is no guarantee that the IP addresses used in the customer VPN are globally unique.

Drivers for IP-Based Virtual Private Networks

Progress marches ever onward, and the world of networking is no different (McDysan, 2000). Similarly to the way enterprises constructed private data networks over the telecommunications infrastructure developed for telephony, the industry is developing a new wave of technologies, overlaying the basic suite of Internet protocols, to construct VPNs. When the public network infrastructure of a VPN matches that of the enterprise equipment, then significant savings can occur. This is a recurring theme in the history of communication networks, with the Internet simply the latest frontier.

Successful enterprises are cost conscious. Even large government programs are subject to public scrutiny. In the highly competitive world of commercial enterprises, those that are not cost conscious fail on a predictable and

regular basis. Standing still is simply not good enough. The maturation of computing hardware and the supporting software has ushered in the postindustrial information age. Now, enterprises need to interconnect employees, databases, servers, affiliates, and suppliers in a rapidly changing business environment. Flexibility becomes an overarching requirement. Those enterprises that do not adapt will not survive.

Increased competition breeds the need for innovation. In traditional services and products, new, smaller companies grab market share by offering new and innovative services more rapidly, or by offering traditional services or products at a lower cost. The incumbents sometimes cry foul, claiming that the newcomers are "cream skimming" the lucrative market segments. The newcomers counter that the incumbents are the "fat cats," who have all the cream. Although some monopolies do exist, either regulated or de facto, the pace of change is ever accelerating.

The worldwide adoption of the Web is a great equalizer. Even a small enterprise can have a large impact and presence via the electronic Web that never sleeps. The user-friendly Web browser with downloadable plug-ins empowers distribution of new paradigm-shifting applications within days to weeks. The rapid adoption of electronic commerce will forever change the way business operates and government administrates. Enterprises are rapidly deploying Web-based intranet and extranet technology to reduce internal costs, in many cases replacing legacy mainframe-based systems.

Communication networks continue to shrink the distances between nations, cultures, and time zones. The introduction of each new type of communication technology empowers the nearly instantaneous dissemination of new media types around the globe. Beginning with the first transatlantic telephone cable in 1956, the speed of transfer of news and breaking information fell from days to minutes. Communications satellites ushered in the era of video and multimedia distribution in the 1960s, on the heels of the space age. In the late 20th century, high-capacity fiber optic transoceanic and transcontinental cables connected the planet, bringing the benefits of digital transmission to the corridors used by most enterprises. This increase in high-performance connectivity enables enterprises to scale beyond national boundaries, particularly in the commercial and nonprofit sector, and it also has an impact on governmental enterprises. Witness the lowering of national barriers in the European Union, as an example.

Most enterprises have some sensitive information that would be of value to competitors or other parties. Enterprises trust the implicit security in private leased-line networks. In fact, a major impediment to the adoption of VPNs is ensuring that this new technology delivers the level of privacy and security that enterprises have come to expect from private lines. Toward this end, the fundamental security requirements of any VPN are the following (Kosiur, 1998; Schneier, 1995; McDysan, 2000): *authentication,* validating that originators are indeed who they claim to be; *access control,* the act of allowing only authorized users admission to the network; *confidentiality,* ensuring that no one can read or copy data transmitted across the network; and *integrity,* guaranteeing that no one can alter data transferred by the network.

VPN approaches employ different methods to meet these requirements. These methods are sometimes implicit and sometimes explicit. Security is a fundamental requirement for customer-edge (CE)-based VPNs operating over the shared Internet infrastructure. Of course, good security begins with secure practices. For example, if the employees of an enterprise leave their user IDs, passwords, or encryption keys lying around, then all the security technology in the world won't protect sensitive information.

Most enterprises believe that quality of service (QoS), traffic management, and prioritized or differentiated service will become an increasingly important driver in their evolving communications needs. Some applications, such as voice and video, require rigid amounts of capacity and minimum levels of quality to operate acceptably. Other applications, such as Web browsing, file transfers, and e-mail, are elastic and can adapt to available capacity to a certain extent. However, even elastic applications result in lowered productivity and increase effective cost to the enterprise if certain minimum-capacity and -quality guidelines are not met. Normally, an enterprise may also need to prioritize or differentiate between these categories of applications to handle intervals of congestion.

The primary QoS measures are loss, delay, jitter, and availability. Voice and video applications have the most stringent delay, jitter, and loss requirements. Interactive data applications such as Web browsing and electronic collaboration have less-stringent delay and loss requirements. Non-real-time applications, such as file transfer, e-mail, and data backup, work acceptably across a wide range of loss rates and delay. Availability requirements vary across enterprises.

Capacity, also referred to as bandwidth, is fundamental to the traffic engineering of a VPN, which is necessary to deliver the required QoS. Some applications require a minimum amount of capacity to work at all, for example voice and video. The performance of elastic protocols that adaptively change their transmission rate in response to congestion in the network improves as the capacity allocated to them increases. The Internet's transmission control protocol (TCP), which carries Web traffic and file transfers, is an example of an elastic protocol. Other applications are elastic up to a certain point, after which adding capacity does not improve performance.

Many network providers guarantee specific QoS and capacity levels via service level agreements (SLAs). An SLA, which is a contract between the enterprise user and the network provider, spells out the capacity provided between points in the network that should be delivered with a specified QoS. If the network provider fails to meet the terms of the SLA, then the user may be entitled to a refund. These have become popular capabilities offered at additional cost by network providers for the private line, frame relay (FR), asynchronous transfer mode (ATM), or Internet infrastructures employed by enterprises to construct VPNs.

Several approaches have been standardized for delivering one or more of the above aspects of QoS. The oldest is the integrated services (Intserv) architecture (RFC 1633,

Braden, Clark, & Shenker, 1994) that uses the resource reservation protocol (RSVP) (RFC 2210, Wroclawski, 1997). Intserv/RSVP allows a host to request one of several levels of QoS at a specified level of capacity for a flow of packets specified by the IP address, transport protocol port numbers, and/or protocol type. The RSVP messages normally follow the same hop-by-hop routed path as other packets, and if the reservation is successful, then the network provides the requested QoS for the level of capacity reserved. However, because RSVP signaling occurs at the individual flow, there is a significant scalability issue in a provider's backbone network due to the signaling load for a large number of flows. For this reason, Intserv/RSVP is not supported in service provider networks and has seen only limited use in enterprise networks.

In responses to these issues, the IETF defined another approach, which addresses the scalability issues of Intserv/RSVP by treating only aggregates of flows using a convention called differentiated services (Diffserv) (RFC 2475, Blake et al., 1998). Diffserv redefines the type-of-service (TOS) byte in the IP packet header in terms of a small number of Diffserv code points (DSCPs), which indicate the type of QoS the packet should receive. Capacity reservation at the individual flow level of Intserv/RSVP is avoided altogether and replaced by classification and traffic conditioning (e.g., policing) performed only at the edge of a DiffServ domain, for example a customer network or a provider network. Furthermore, because Diffserv operates only on fields within the IP packet header, it can coexist with IP security protocols whereas Intserv/RSVP may not, because it may rely on higher-layer protocol fields (e.g., transport protocol port numbers) to identify an individual flow.

Most backbone IP networks will likely use DiffServ, possibly using a so-called bandwidth broker, which incorporates policy server functions and also deals with customer traffic contract and network resource allocation. A bandwidth broker maps service level specifications to concrete configurations of edge routers of a DiffServ domain. However, Intserv/RSVP or next-generation reservation signaling protocols still might have a role to play in signaling reservations in enterprise networks and at the edge of a service provider network, especially for such applications as digital audio and video, which would benefit from reservations for relative long-lived, high-bandwidth flows (Braun, 2001).

Introduction to Virtual Private Networks Technologies

A VPN attempts to draw from the best of both the public and the private networking worlds. Such a network is private in the sense that the data an enterprise transfers over the VPN is separated and/or secure from that of other enterprises or the public. It is virtual in the sense that the underlying public infrastructure is partitioned to have some level of service for each enterprise. A VPN is communication between a set of sites making use of a shared network infrastructure, in contrast to a private network, which has dedicated facilities connecting the set of sites in an enterprise. To a great extent, the intent is that the logical structure of the VPN, such as topology, addressing, connectivity, reachability, and access control, is equivalent to part or all of a conventional private network.

A good VPN has the low-cost structure of a ubiquitous public network but retains the capacity guarantees, quality, control, and security of a private network. How can a network design achieve these apparently contradictory goals? The answer lies in software-defined networking technology, sophisticated communications protocols, and good old-fashioned capitalism.

FR, ATM, multiprotocol label switching (MPLS), and the Ethernet are all forms of layer 2 (L2) label-switching protocols (McDysan, 2000). A label is the header field of a packet, frame, or cell. Labels are unique only to an interface on a device, such as enterprise user equipment or a network switch. Figure 1 illustrates a simple example of the operation of a simple two-port label switch. A label switch uses the label header from the packet received on an interface (left side of figure) as an index into a lookup table in the column marked "In," which identifies a specific row. From this row, the lookup table returns the outgoing label from the column marked "Out" and the outgoing physical interface from the column marked "Port."

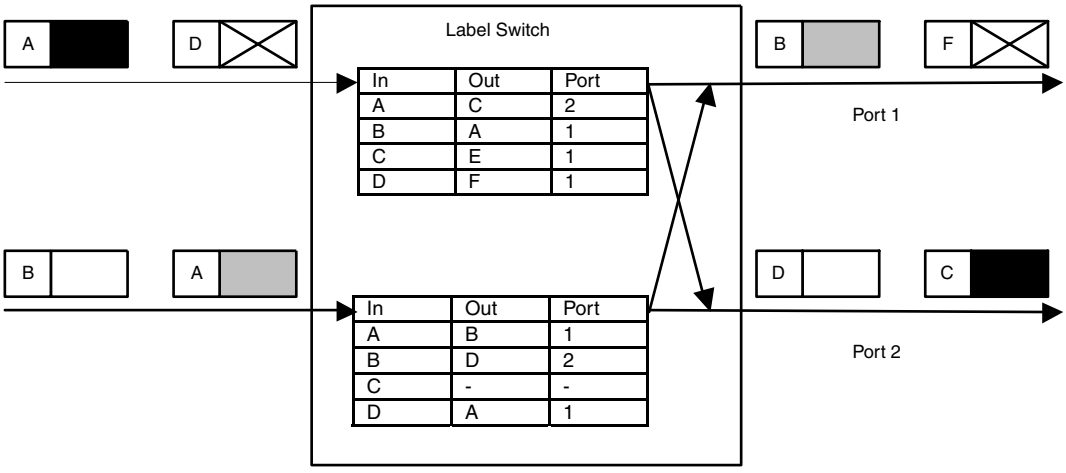


Figure 1: Illustration of layer 2 label switching.

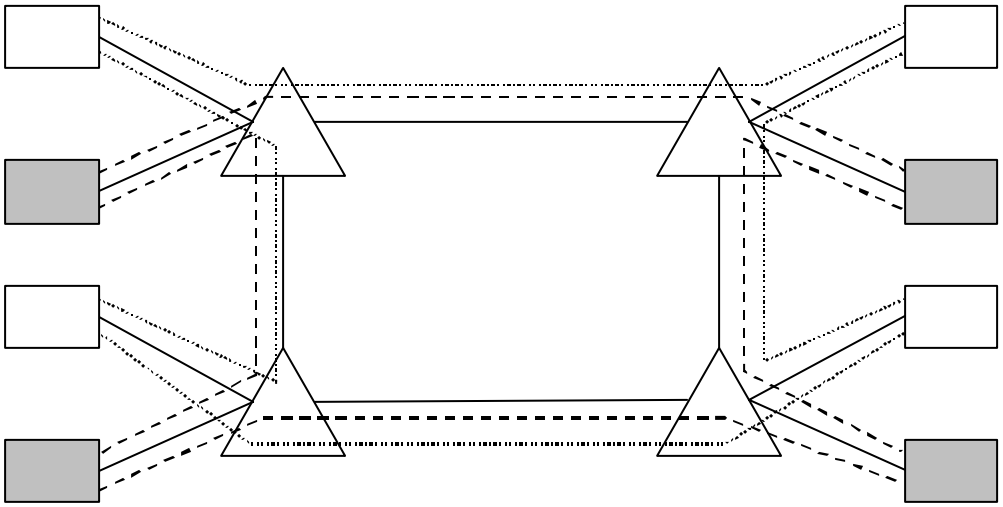


Figure 2: Example of two connection-oriented VPNs in a shared public network.

The switch routes the packet, frame, or cell to the outgoing physical interface using an internal switching fabric and "switches" the label to the outgoing label retrieved from the lookup table. The example in the figure uses patterns for the packets to trace the result of the label-switching operation implemented by the lookup tables on the input side of each port. Of course, contention may occur for the output port in a label switch if multiple packets are destined for the same output. Typically, label switches must implement some form of queuing to handle this situation.

An L2 network consists of a number of label switches implementing the basic function described above. Typically, these switches also implement a number of other features related to connection establishment, traffic control, QoS, congestion control, and the like. Some form of routing, signaling, and/or network management protocol establishes a consistent sequence of label-switching mappings in the lookup tables to form a logical connection that can traverse multiple nodes. When the network is connection oriented, for example in FR, ATM, and MPLS, we call the allowed pairwise communication a virtual circuit or connection (VC). For a connectionless L2 network, such as the Ethernet, we call the set of sites that are allowed to communicate a virtual local area network.

Figure 2 illustrates a public connection-oriented network supporting two disjoint VPNs. Shaded boxes represent equipment from different enterprises at various sites connected to triangles that represent provider-edge (PE) label switches. The label-switched connection-oriented network implements disjoint virtual connections (either permanent or switched) between different enterprise nodes, as indicated by dashed lines of different styles in the figure. A connection-oriented label-switched network operates very much like a private line network, but it uses virtual connections instead of real ones. The important difference is that the service provider switches utilize label switching instead of Time Division Multiplexing (TDM) cross-connects to logically share trunk circuits between multiple enterprise VPNs. Thus, a connection-oriented VPN can be a plug-compatible replacement for a private-line-based network. This has a number of advantages. First, the granularity of capacity allocation is much finer with a label switch than with that implemented in the rigid TDM hierarchy. Second, if the traffic offered by the enterprises is bursty in nature, the service provider network can efficiently multiplex many traffic streams together. Finally, the shared public network achieves economies of scale by utilizing high-speed trunk circuits that have a markedly lower cost per bit per second (bps) than lower-speed links do.

X.25 was the first connection-oriented data VPN, but it is now being phased out. X.25 pioneered a VPN concept called a closed user group (CUG), which is similar to that of an intranet or extranet. In the late 1980s, FR followed X.25, simplifying the protocol and, hence, improving the price-performance ratio. FR pioneered the important VPN concept of per-connection traffic management and some simple responses to congestion. ATM was the successor to FR, in the mid-1990s, focusing on a fixed cell size to ease hardware implementation and achieve high performance. ATM borrows heavily from the signaling protocols of the narrowband integrated services digital network (ISDN), the traffic management concepts of FR, and automatic topology discovery from IP. ATM standards significantly extended the concept of QoS and more precisely defined traffic management, these being the hallmarks of ATM. In some ways, MPLS is an enhancement of ATM: It provides most of the same capabilities but also adds some useful extensions and refinements tailored to the support of IP. MPLS overcomes the inefficiency caused by the partial fill of the last fixed-length ATM cell when carrying variable-length packets in AAL5. MPLS also supports a more flexible hierarchical aggregation of connections and supports loop detection as well. The design of MPLS also allows tighter integration than did ATM of connection-oriented traffic engineering with IP routing protocols in service provider backbones. Extensions of these capabilities are also quite useful in support of network-based VPNs.

A connectionless protocol like IP does not require a signaling protocol because it does not use connections

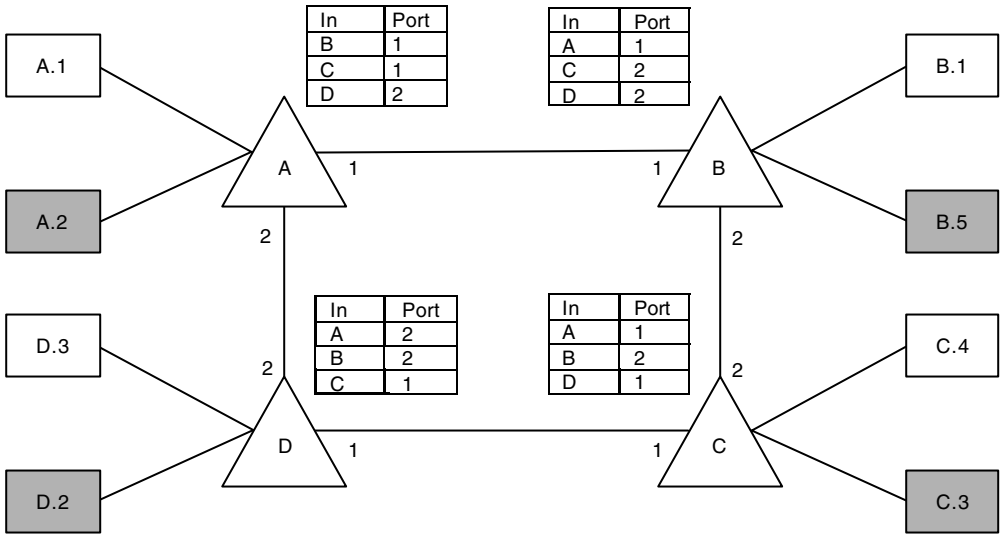


Figure 3: Example of two connectionless VPNs.

to forward user traffic. Instead, a routing protocol distributes topology information such that each node can make an independent, yet coordinated, decision about the next hop on which to forward packets that have a particular destination address prefix in the header. Unlike label switching, the addresses in packet headers must be unique throughout a set of interconnected networks, such as the Internet. Therefore, the forwarding lookup table is identical in every node in a simple connectionless network. Because each address must be unique, the forwarding table could become quite large. The Internet scales to large sizes by carefully administering address assignments so that their forwarding tables need only process the high-order prefix bits of the address.

In a connectionless network, a VPN is a logical overlay on a shared IP network of a different type. A shared IP network may be the public Internet or a network that supports IP routing protocols implemented specifically for use by enterprise customers. A secure IP VPN utilizes the concept of an encrypted tunnel implemented at the enterprise equipment connected to the IP network. A tunnel may exist at the link layer or the network layer as an association between two endpoints attached to a public network, therefore making it virtual. Encryption is a technique that scrambles information such that only the intended receiver can decode it, thereby achieving privacy. Because an IP network is connectionless, the packets between enterprise nodes may take different paths, depending on such conditions as link failures or the configuration of routing parameters. IP routing protocols synchronize the forwarding tables in all the nodes whenever the state of the network changes. This fundamental difference in paradigms is what has allowed the Internet to scale the way it has in response to the tremendous demand that arose in the latter half of the 1990s.

Figure 3 illustrates a connectionless IP-based VPN for two enterprises. The enterprise nodes are shaded boxes, each with an IP address that has a prefix (e.g., A.1 and B.5) associated with a triangle indicating the network router to which the access line attaches (e.g., A and B). For example, the gray-shaded enterprise node has an address prefix A.2 connected to the network router with address prefix A. The figure illustrates the forwarding tables next to each network router. Each table contains an entry labeled "In" for the incoming packet address prefix, which is used to look up the next-hop outgoing port. For example, at router A, a packet received with destination address prefix B is sent out on port 1. Note how these tables contain only the address prefix and the next-hop link number, and not the enterprise node address prefixes. Therefore, the enterprise equipment at the edge of the network implements the IP VPN functions. This architecture has a number of fundamental advantages. First, configuration changes to the enterprise VPN do not require changes in the core Internet. Second, because the Internet is a global public network, a tunneled enterprise VPN can be implemented across multiple Internet service provider (ISP) networks.

Now we look at a categorization of logical VPN types and the terminology used to describe them.

A Taxonomy of IP-Based Virtual Private Networks

The taxonomy of VPN types is primarily determined by whether the tunnels that provide the service terminate on CE or PE devices (Carugi et al., 2002; Callon et al., 2002). Figure 4 illustrates the case where the tunnels terminate on the CE. A CE-based VPN is one in which knowledge of the service aspects of the customer network is limited to CE devices. Customer sites are interconnected via tunnels or hierarchical tunnels, as defined in the glossary. The service provider network is unaware of the existence of the VPN because it operates exclusively on the headers of the tunneled packets. Specifically, a CE-based L2 VPN is a link layer (i.e., L2) service provided by CE equipment at the customer sites, for example the Ethernet. In a similar manner, a CE-based L3 VPN is a network layer (i.e., L3) service provided by CE devices at customer sites, for example the IP.

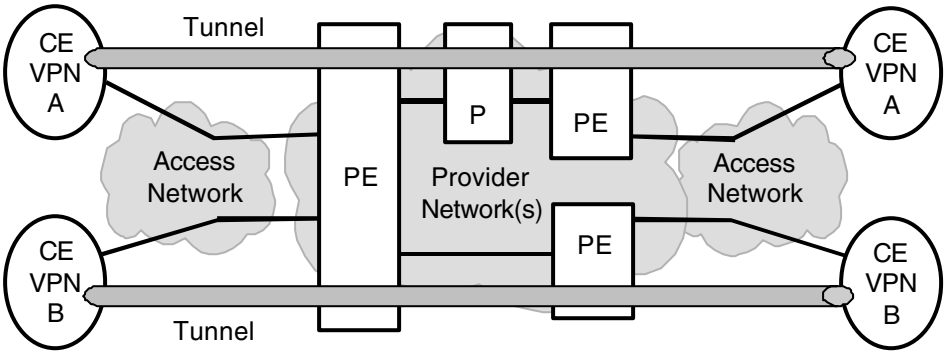


Figure 4: Generic customer edge (CE)-based VPN.

Figure 5 illustrates the case where the tunnels terminate on the PE. A PE-based VPN is one in which the service provider network maintains state information for each customer VPN such that packets are forwarded between customer sites in an intranet or extranet context using the customer's address space. Often, a hierarchical tunnel is used between PEs, with the outermost tunnel being implemented by a provider (P) router, which provides PE–PE connectivity. (Note that the P and PE functions are logical and that a single router may implement both functions.) These tunnels may be dedicated to separate VPNs or they may be shared between multiple VPNs by the PEs, which use label stacking to isolate traffic between VPNs. These inner tunnels interconnect an L3 virtual forwarding (or L2 switching) instance (VFI/VSI) for each VPN instance in a PE switching router. A PE-based L2 VPN provides an L2 service that switches link-layer packets between customer sites using the customer's link-layer identifiers, for example the Ethernet. A PE-based L3 VPN provides an L3 service that routes packets between customer sites using the customer network's address space, for example the IP.

The CE-based approach is the simplest from the service provider backbone perspective, but it requires a fair amount of configuration and management of the CE. On the other hand, the network-based approach provides greater control of traffic engineering and performance, but it incurs additional complexity in the backbone network to achieve these benefits. The L3 PPVPN framework document (Callon et al., 2002) further describes these concepts in the context of a reference model that defines layered service relationships between devices and one or more levels of tunnels. The next sections cover some specifics of CE- and PE-based VPNs as they relate to IP intranets and extranets.

CUSTOMER-EDGE-BASED VIRTUAL PRIVATE NETWORKS

As defined earlier, CE-based VPNs are partitioned by tunnels established between CE devices. Routing inside the customer network often treats the tunnels as simple point-to-point links, or sometimes as broadcast local area networks. For customer-provisioned CE-based VPNs, provisioning and management of the tunnels is up to the customer network administration, which is also responsible for operation of the routing protocol between CE devices. In provider-provisioned CE-based VPNs, the service provider(s) perform provisioning and management of the tunnels and may also configure and operate routing protocols on the CE devices. Of course, routing within a site is always under control of the customer.

There are two primary types of IP CE-based VPNs, distinguished by the type of tunnel employed. The first is older and is used primarily to construct intranets by using CE routers connected via FR or ATM virtual connections. The second is newer and is based upon tunnels implemented using cryptographic methods over the public Internet using either dedicated or dial-up access. We now describe each of these approaches.

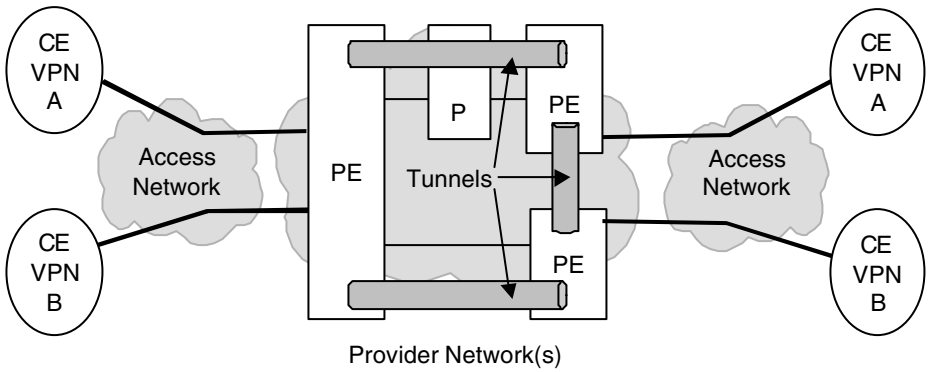


Figure 5: Generic PE-based (also called network based) VPN.

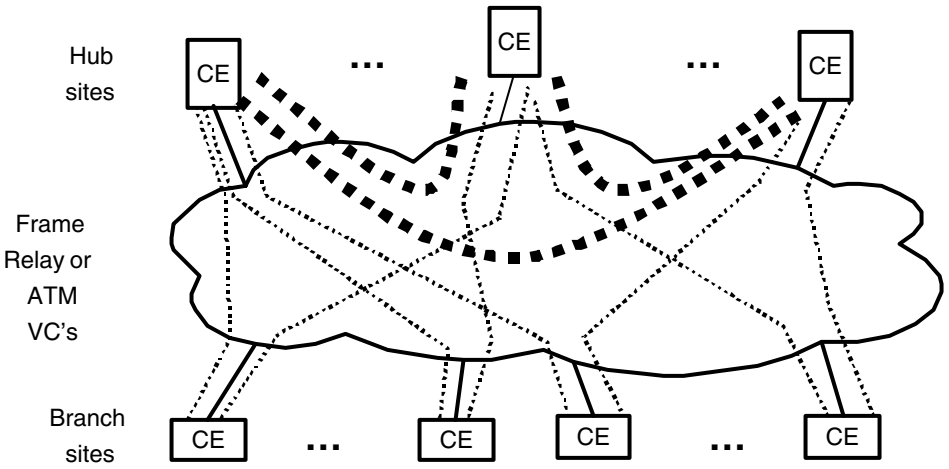


Figure 6: CE-based VPN over a partial mesh of L2 hub-and-spoke VCs.

CE Virtual Private Networks Over Virtual Connection Networks

The FR and ATM connection-oriented VPN alternatives largely apply to a single service provider. In order to connect each site to every other site in a fully meshed network of *N* number of sites, the service provider must provision on the order of *N* squared virtual connections (VCs). Note that each VC must be provisioned at every intermediate FR or ATM switch in the service provider network. As the number of sites becomes large, service providers often interconnect the sites, creating what is called hub-and-spoke architecture, as shown in Figure 6. Often, the hub sites are connected in a full mesh with branch sites dual-homed to a primary and secondary hub site, as shown in the figure. Another motivation for the hub-and-spoke design is that with a full mesh of sites, addition of a new site requires configuration not only of the new site but of each of the other VPN sites as well.

The traffic forwarded between the sites in a VPN is isolated from all others by the logical separation provided by the virtual connections, which perform label switching as configured by a provisioning system. What results is, for all practical purposes, a private network. Such a connection-oriented VPN is a good approach for intranets because of the isolation and site-to-site traffic engineering, provided by the approach is good.

On the other hand, configuring such a network for extranets can be complex and inflexible. For these reasons, e-commerce applications tend to use IP security protocols as the foundation for CE-based VPNs that are used by many intranet and extranet applications.

IP Security-Based Customer-Edge Virtual Private Networks

An analogous IP-based VPN network has the same number of hub-and-spoke sites but requires the addition of overlay IP security (IPsec) tunneling and/or encryption functions in the CE devices. There is no explicit connection through the devices in the service provider network. Instead, all the tunnel functions are implemented in the CE devices. Scaling issues similar to those in CE devices overlaid on virtual connections arise in IPsec CE-based VPNs, but here the limits are the number of IPsec tunnels and the number of routing adjacencies a CE router can support. Therefore, large IPsec CE-based VPNs also have a hub-and-spoke architecture, as described previously. Figure 7 illustrates the same hub-and-spoke network example, with circles showing the hub–spoke tunnels and

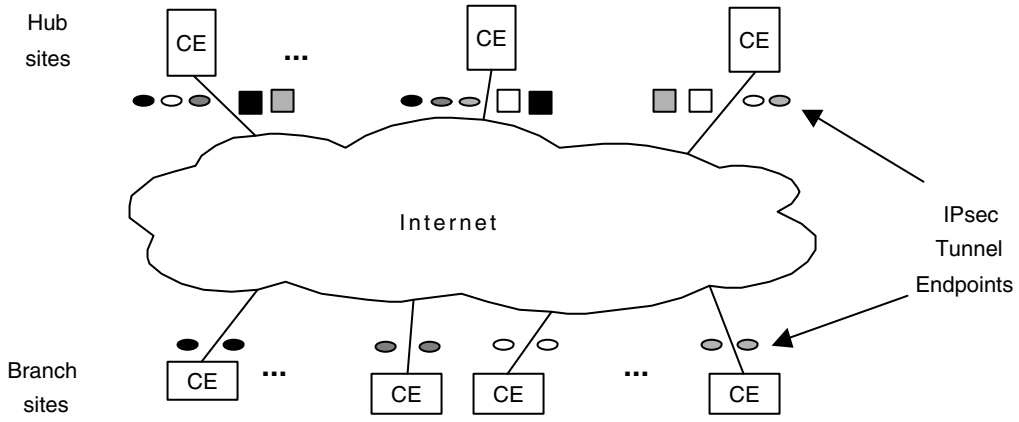


Figure 7: CE-based VPN using IPsec tunnels over the Internet.

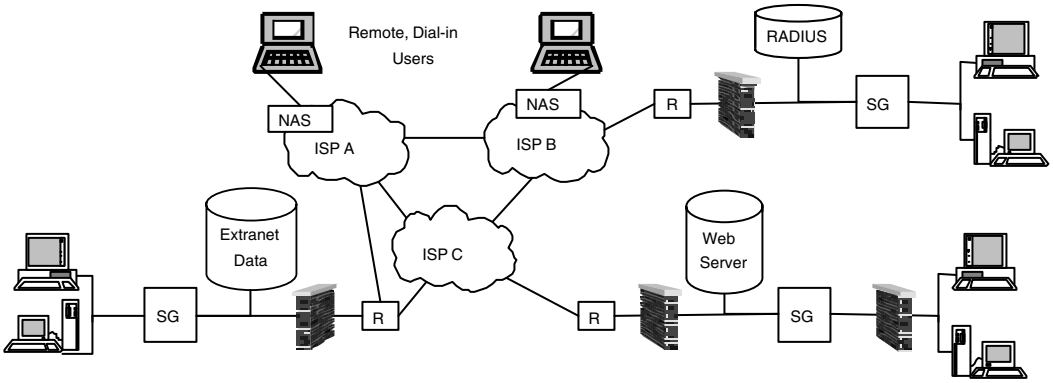


Figure 8: Pure IP-distributed VPN design.

squares showing the hub–hub tunnels. As with the VC overlay approach, adding a new site to a full mesh requires configuration of a tunnel to every other site. Furthermore, if the enterprise does not use globally unique, routable IP addresses, the CE devices may also include network address translation functions. When a single ISP provides the network for an IP-based VPN, then guarantees on quality and performance are feasible. Beware of an IP-based VPN built on top of the public Internet using services provided by several ISPs: It may not provide the quality necessary for telephone-grade voice or multimedia applications.

The IETF designed the IPsec protocol suite to address the known issues involved with achieving secure communications over the Internet (McDysan, 2000). It reduces the threat of attacks based on IP address spoofing and provides a standardized means for ensuring data integrity, authenticating a data source, and guaranteeing confidentiality of information. Furthermore, it tackles the complex problem of key management head on. When a public key management infrastructure is used, the Internet can be trusted based upon this set of standards. IPsec will play an important role not only in enterprise VPNs, but also in electronic commerce and in secure individual end user communication.

IPsec refers to a suite of three interrelated security protocols implemented by modification to, or augmentation of, an IP packet in conjunction with an infrastructure that supports key distribution and management. An interrelated set of Request for Comments (RFCs) published by the IETF specifies the details of IPsec. RFC 2401 (Kent and Atkinson, 1998) describes the overall IP security architecture, whereas RFC 2411 (Thayer et al, 1998) gives an overview of the IPsec protocol suite and the documents that describe it. Three protocols make up IPsec, with the names identifying the function performed. The two primary protocols involved in the transfer of data are called the authentication header (AH) and the encapsulating security payload (ESP). The AH protocol provides source authentication and data integrity verification using a header field, but it does not provide confidentiality. AH also supports an optional mechanism to prevent replay attacks. The ESP protocol uses both a header and a trailer field to provide confidentiality via encryption. ESP may also provide data integrity verification, source authentication, and an antireplay service. Because both the AH and the ESP protocols utilize cryptographic methods, secure distribution and management of keys is a fundamental requirement. IPsec specifies that key management may be manual or automatic. The automatic key management protocol specified for IPsec is called Internet key exchange and involves the mechanism for creating a security association (SA) between a source and a destination for the AH and ESP protocols.

The AH and ESP protocols operate in either transport or tunnel mode, as defined by the parameters of an SA. In *transport mode,* they provide security by creating components of the IPsec header at the same time the source generates other IP header information. This means that transport mode can operate only between host systems. In *tunnel mode,* IPsec creates a new IP packet, which contains the IPsec components and encapsulates the original unsecured packet. Because tunnel mode does not modify the original packet contents, it can be implemented using hardware or software located at an intermediate security gateway (SG) between the source or destination system.

Figure 8 illustrates a pure IP-based VPN design that has a cost structure essentially independent of the traffic pattern. Here, every site has a firewall and security gateway, so any site may directly access the Internet or any other site. In addition, we show a network access server (NAS), remote authentication dial-in user service (RADIUS) server, Web server, and extranet database located at three separate sites. Dial-in users are secured using the RADIUS server and the SG. This design also reduces access costs because traffic for the Internet need not traverse a firewall at a headquarters site, as shown in the hierarchical example above. Sites may also be dual-homed to different ISPs or to different sites within the same ISP for resiliency purposes, as necessary. This design is better suited to extranet applications and electronic commerce because communication via the public Internet is more interoperable and rapidly deployable than any other communication service.

PROVIDER-EDGE-BASED LAYER 3 VIRTUAL PRIVATE NETWORKS

A PE-based VPN is one in which PE devices in the service provider network provide the partitioning of forwarding and routing information to only those (parts of) sites that are members of a specific intranet or extranet. This allows

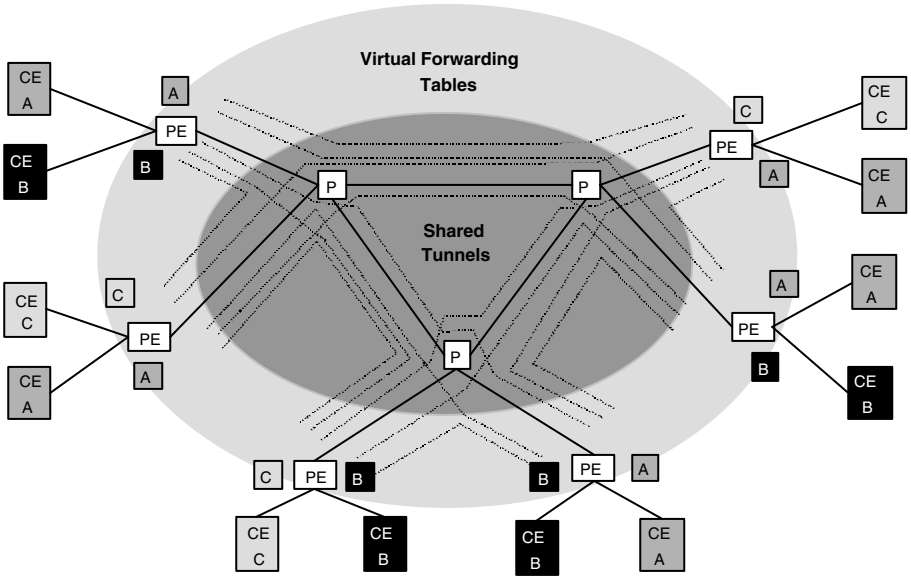


Figure 9: Aggregated routing and shared tunnel network-based L3 VPN.

the existence of the VPN to be hidden from the CE devices, which can operate as though they were part of a normal customer network. As described earlier, PE-based VPNs use tunnels set up between PE devices. These tunnels may use one of a number of encapsulations to send traffic over the provider network(s), for example MPLS, generic routing encapsulation, IPsec, or IP-in-IP. As sites for new VPNs are added or removed, PE-based VPN solutions provide a means of distributing membership information automatically. There are two principal methods defined in the IETF (Callon et al., 2002) for implementing these types of PE-based VPNs, namely aggregated routing and virtual routers, which we now describe.

Aggregated Routing Virtual Private Networks

The aggregated routing approach is one in which a separate forwarding table exists for each VPN on every PE that connects to a site in that VPN but where the exchange of routing information between the PEs is multiplexed, or aggregated together. The BGP/MPLS VPN (RFC 2547, Rosen & Rekhter, 1999) approach uses extensions to the border gateway protocol (BGP) to implement this generic architecture. Figure 9 illustrates an example of this approach, connecting sites from three VPNs, A, B, and C, in an extranet. Each PE has a separate virtual forwarding table for each VPN site that it serves, but the forwarded traffic and exchanged routing information uses a set of shared tunnels, as shown in the center of the figure. Often these types of solutions are implemented on a single service provider network. However, there are some implementations across more than one provider network.

This approach alleviates some of the scaling issues involved with the connection- or tunnel-oriented CE-based approaches described earlier when full communication between a set of sites is desired. Specifically, when adding or removing a site, only the PE involved with that site need be reconfigured—the BGP/MPLS protocols automatically take care of the rest. Furthermore, the protocols have the capability of advertising to their peers more than one route for the same destination address. This can be useful in an extranet to force traffic exchanged between different enterprises through additional devices, such as firewalls or filters.

Virtual Router Virtual Private Networks

Although the virtual router (VR)-based approach (RFC 2917, Muthukrishnan & Malis, 2000) also uses PE and P routers, there are several important differences, as illustrated in Figure 10. This example uses the same CE sites from the three VPNs discussed in the aggregated routing example above. In a VR VPN, a VR is dedicated to each VPN in every PE that supports a site for that VPN. This means that each enterprise can manage its own routing on the VR in the PE. This works very well in cases where the enterprise network has other forms of connectivity between its sites: The VRs look like just another (well connected) router to the enterprise network. Usually, a separate set of tunnels is allocated in a full mesh between the VRs, as shown by different line styles in the center of the figure. This allows excellent control of capacity allocation and control of QoS between the VPN sites.

The VR PE-based VPN is best suited for intranets. It is not frequently used in an extranet because one enterprise would have to exchange routing information with another. This could lead to undesirable security holes, instability of the routing, and, hence, a greater likelihood of an outage, as well as more difficult coordination in the event of the inevitable moves, adds, and changes. It could be used, however, as a backbone network provided by one partner for connecting a number of other enterprises together, for example using CE-based VPNs overlaid on a managed VR PE-based network.

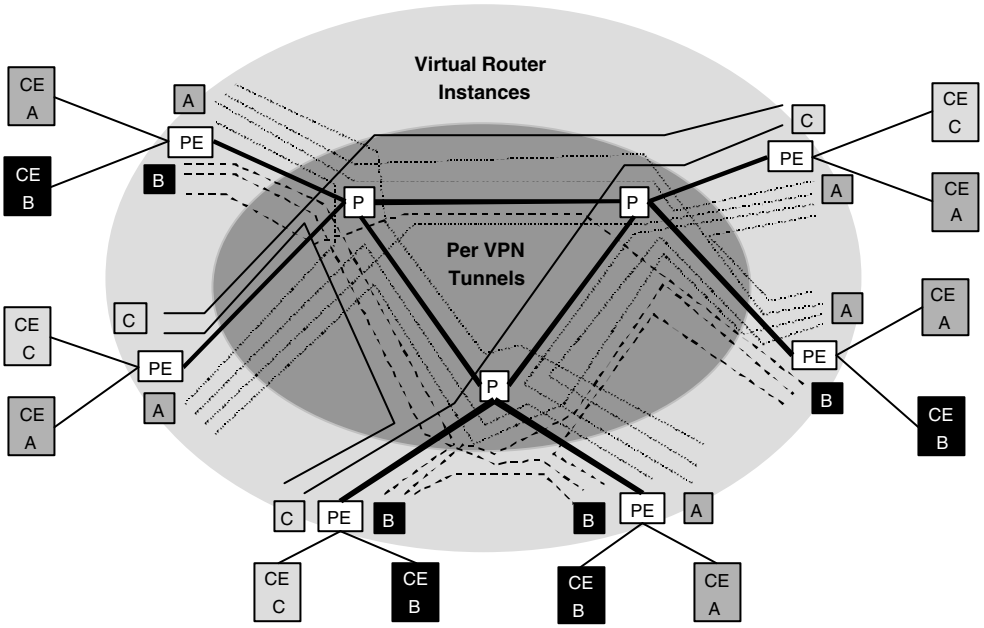


Figure 10: PE-based L3 VPN with virtual routers with tunnels per VPN.

DESIGN CONSIDERATIONS AND EXAMPLE OF VIRTUAL PRIVATE NETWORKS

This section summarizes some important considerations when choosing a VPN approach and gives an example of a CE-based IPsec VPN used for electronic commerce.

Considerations When Choosing a Virtual Private Networks Approach

Establishing a set of goals and establishing a plan to meet them is critical to success in most human endeavors, and virtual private networking is no exception (McDysan, 2000). The steps here are similar to that of any large-scale project. First, researching requirements, drivers, and needs is necessary to establish goals. Next, developing several candidate designs and analyzing them in the harsh light of commercial business reality is a crucial step. A VPN may not be right for the enterprise under consideration at this time, and timing is important. Finally, a decision to implement a new type of VPN or to migrate existing private network applications to a VPN, is but the first step of many. Detailed planning and a well thought out migration strategy are essential for an enterprise to achieve its goals identified in the first step above.

A number of enterprises have already implemented VPNs of the types described in this chapter. A good starting point is to look at an enterprise that is similar to yours in some way and to read case studies, papers, and books about what worked and what did not. However, be aware that the needs of each enterprise are unique, and therefore basing a decision upon another's experiences, while helpful, cannot guarantee that goals will be met.

An important area of requirements research is analyses of potential security threats and essential performance metrics. Formulating a threat model and considering what would happen if important information were stolen, made public, or corrupted is an essential step. Determining the performance required by applications is also important. Consider what would happen if a site were disconnected for a long period of time. Assess what the impact of network congestion would be. Discriminate between what would be nice to have and what is absolutely necessary in the way of performance—this can make quite a difference in qualifying network designs and their eventual cost.

Although a generic framework may not apply to all enterprises, there are some helpful points to consider when categorizing types of requirements. One way to analyze VPN requirements is to consider the community of interest and the access methods: cost-effective remote and mobile user access; an infrastructure for intranets that keeps resources secure within a single enterprise; an infrastructure for extranets for controlling resource sharing between two or more enterprises.

The economic crossover point regarding enterprise dial-in versus ISP-provided access services centers around the number of users that require dial-in access and the type as well as amount of activities these users conduct. In general, a remote user population that generates bursty activity during relatively long duration sessions is a good candidate for ISP access. As described earlier, most VPN techniques differ in the degree of traffic separation and control that an enterprise can have in an intranet context. On the other hand, if a driving requirement for the enterprise is extranet connectivity, then an IPsec-based solution is one of the few choices available (for more information, see VPN Consortium, 2003).

Because this is such an important case in the world of electronic commerce, we now look at an example where a few large enterprises worked with a number of small-to-medium-size enterprises to create a successful model for extranet deployment.

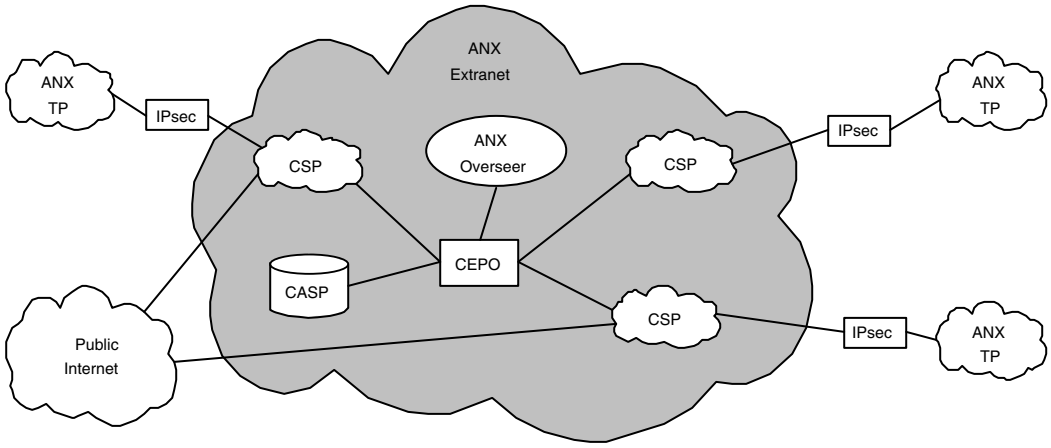


Figure 11: ANX extranet architecture.

Example of Deployment of a Customer-Edge-Based Virtual Private Networks in E-commerce

Unless your enterprise is the first to try a new technology, protocol, or architecture, there will likely be case studies available for review. A frequently documented extranet case study is the Automotive Network eXchange (ANX) (McDysan, 2000). This extranet VPN involves a few large enterprises (automotive manufacturers) and a significant number of small-to-medium-size enterprises (their suppliers). Initiated by the Automotive Industry Action Group (AIAG) in 1994, the IPsec-based ANX network had Chrysler, Ford, and General Motors as the founding network participants. These companies and other major automotive manufacturers utilize parts and services from a large number of common original equipment manufacturers, such as Bosch, Delta, Fisher, ITT, and TRW. Following the completion of successful trials in 1997 and 1998, ANX launched production in November 1998. By the end of 1999, ANX had nearly 500 registered trading partners. As an example of a quantifiable goal achievable by an extranet, the AIAG estimates that a collaborative planning, forecasting, and replacement tool running over the ANX network may save up to $1,200 per vehicle. This savings results from a reduction of the delivery cycle of parts and supplies and the associated inventory levels.

The ANX architecture is based upon a set of interconnected certified service providers (CSPs), certified exchange point operators (CEPOs), and certificate authority service providers to which ANX trading partners subscribe, as illustrated in Figure 11. Telcordia (formerly Bellcore) has been chosen as the ANX overseer, which awards certification to CSPs and CEPOs. The ANX service quality certification categories are network service features, interoperability, performance, reliability, business continuity and disaster recovery, security, customer care, and trouble handling. ANX has also specified that the International Computer Security Association (ICSA) will certify whether equipment is IPsec compliant.

Finding companies with equipment that has the ICSA stamp of approval is a good place to start when looking for IPsec-compliant vendors.

This network is effectively a partitioned set of interfaces running on top of the public Internet infrastructure offered by the selected set of certified commercial ISPs. It replaces the prior complex arrangement of physical and logical connections between trading partners with one logically administered, cryptographically secured connection to the ANX extranet. Choice of the TCP/IP protocol suite provides access to a broad range of file transfer, electronic document interchange, e-mail, and other application software. This is especially important in the automotive industry, where computer-based techniques are now used in almost every stage of the design, manufacturing, delivery, and maintenance aspects of the business. Although the benefits of ANX apply primarily to medium-to-large-size enterprises in the automotive industry, the drive toward interoperability will benefit other industry segments in the longer term (for more information, see www.anx.com).

GLOSSARY

Customer-edge (CE) device Provides access for users at a site and has an access connection to a PE device. It allows users at a site to communicate over the access network with other sites in the VPN.

Enterprise A single organization, corporation, or government agency that administratively controls and sets policy for communication among a set of sites.

Extranet Allows communication between a set of sites that belong to different enterprises, as controlled by the enterprise administrators and/or a third party. These enterprises have access to a specified subset of each other's sites. Examples of extranets include (a) companies performing joint software development, (b) a group of suppliers and their customers exchanging orders and delivery tracking information, and (c) different organizations participating in a consortium that has access to important information.

Generic routing encapsulation (GRE) A general protocol for encapsulating a network layer protocol over another network layer protocol (RFC 2784, Farinacci, Li, Hanks, Meyer, & Traina, 2000).

Intranet Restricts communication to a set of sites that belong to one enterprise and via policy may further restrict communication between groups within these sites. For example, communication between marketing and engineering may be limited.

IP security protocol (IPsec) A set of IETF standards that defines a suite of security protocols that provide confidentiality, integrity, and authentication services (RFC 2401, Kent & Atkinson, 1998).

Layer 2 tunneling protocol (L2TP) An IETF standardized protocol defined initially for support of dial-in connections (RFC 2661, Townsley, et al., 1999). A successor to the proprietary Microsoft PPTP and Cisco L2F protocols, L2TP gives mobile users the appearance of being on an enterprise LAN.

Multiprotocol label switching (MPLS) A switching technique that forwards packets based upon a fixed-length label inserted between the link and network layer or that uses a native layer 2 label, such as FR or ATM (RFC 3031, Rosen, Viswanathan, & Callon, 2001). Similar to frame relay and ATM in function, MPLS differs from these protocols by virtue of its tight coupling to IP routing protocols.

Provider-edge (PE) device A PE device faces the service provider core network on one side and interfaces via an access network to one or more CE devices.

Site A set of users who have connectivity without use of a service provider network, for example users who are part of the same enterprise in a building or on a campus.

Tunnel Formed by encapsulating packets with a header used to forward the encapsulated payload to the tunnel end point. In VPN applications, tunnel end points may be a CE or a PE device. Encapsulating one tunnel within another forms a hierarchical tunnel, which is useful for reducing the number of tunnels in the core of networks. Examples of protocols commonly used for forming a tunnel are MPLS, L2TP, GRE, IPsec, and IP-in-IP tunnels.

User Someone or something that has been authorized to use a VPN service, for example a human being using a host or a server.

Virtual private network (VPN) A specific set of sites configured as either an intranet or an extranet to allow communication. A set of users at a site may be a member of one or many VPNs.

CROSS REFERENCES

See *Circuit, Message, and Packet Switching; Electronic Commerce and Electronic Business; Extranets; Internet Architecture; Internet Literacy; Internet Security Standards; Intranets; Public Networks; TCP/IP Suite.*

REFERENCES

ANX Network (2003). Retrieved February 10, 2003, from http://www.anx.com/

Blake, S., Black, D., Carlson, M., Davies, E., Wang, Z., & Weiss, W. (1998). *An architecture for differentiated services*. Retrieved February 20, 2003, from http://ietf.org/rfc/rfc2475.txt

Braden, R., Clark, D., & Shenker, S. (1994). *Integrated services in the Internet architecture: An overview*. Retrieved February 20, 2003, from http://ietf.org/rfc/rfc1633.txt

Braun, T., Guenter, M., & Khalil, I. (2001, May). Management of quality of service enabled VPNs. *IEEE Communications Magazine*.

Callon, R., Suzuki, M., DeClerq, J., Gleeson, B., Malis, A., Muthukrishnan, K., RosenE., Sargor, C., & Yu, J. (2002). *A framework for layer 3 provider provisioned virtual private networks*. Unpublished manuscript.

Carugi, M., McDysan, D., Fang, L., Nagarajan, A., Sumimoto, J., & Wilder, R. (2002). *Service requirements for provider provisioned virtual private networks*. Manuscript in preparation.

E-mail list logs, presentations, related ITU-T drafts (2003). Retrieved February 20, 2003, from http://ppvpn.francetelecom.com

Farinacci, D., Li, T., Hanks, S., Meyer, D., & Traina, P. (2000). *Generic routing encapsulation (GRE)*. Retrieved February 20, 2003, from http://ietf.org/rfc/rfc2784.txt

IETF working group charter page, list of RFCs and current drafts (2003). Retrieved February 20, 2003, from http://ietf.org/html.charters/ppvpn-charter.html

Kent, S., & Atkinson, R. (1998). *Security architecture for the Internet protocol*. Retrieved February 20, 2003, from http://ietf.org/rfc/rfc2401.txt

Kosiur, D. (1998). *Building and managing virtual private networks*. New York: Wiley.

McDysan, D. (2000). *VPN applications guide*. New York: Wiley.

Muthukrishnan, K., & Malis, A. (2000). *A Core MPLS IP VPN architecture*. Retrieved February 20, 2003, from http://ietf.org/rfc/rfc2917.txt

Rosen, E., & Rekhter, Y. (1999). *BGP/MPLS VPNs*. Retrieved February 20, 2003, from http://ietf.org/rfc/rfc2547.txt

Rosen, E., Viswanathan, A., & Callon, R. (2001). *Multiprotocol label switching architecture*. Retrieved February 20, 2003, from http://ietf.org/rfc/rfc3031.txt

Schneier, B. (1995). *Applied cryptography: Protocols, algorithms, and source code in C*. New York: Wiley.

Thayer, W., Doraswamy N., & Glenn, R. (1998). *IP Security Document Roadmap*. Retrieved February 20, 2003, from http://ietf.org/rfc/rfc2411.txt

Townsley, W., Valencia, A., Rubens, A., Pall, G., Zorn, G., & Palter, B. (1999). *Layer two tunneling protocol L2TP*. Retrieved February 20, 2003, from http://ietf.org/rfc/rfc2661.txt

Wroclawski, J. (1997). *The use of RSVP with IETF integrated services*. Retrieved February 20, 2003, from http://ietf.org/rfc/rfc2210.txt

Virtual Private Network Consortium (2003). Retrieved February 20, 2003, from http://www.vpnc.org/

Virtual Reality on the Internet: Collaborative Virtual Reality

Andrew Johnson, *University of Illinois at Chicago*
Jason Leigh, *University of Illinois at Chicago*

INTRODUCTION

Collaborative virtual reality—sharing immersive computer-generated environments over the high-speed networks—is a next-generation interface that will allow collaborators on different continents to share a space where they can interact with each other and with the focus of their collaboration. This text describes ongoing work in this area at the Electronic Visualization Laboratory at the University of Illinois at Chicago. We first discuss what we mean by the term virtual reality and what the focus is of our work in collaborative virtual environments. We then discuss the types on information that must be sent through the networks to maintain these collaborations. Finally, we describe current research in the areas of asynchronous collaboration and heterogeneous perspectives and conclude with a discussion of what we see as the future of collaborative virtual environments.

VIRTUAL REALITY

Before we discuss collaborative virtual reality, we should define what we mean by virtual reality. Different disciplines have different definitions for what virtual reality is and what hardware is required. A good novel is a form of virtual reality that requires no special hardware to be experienced. For our purposes, virtual reality requires computer-generated stereo visuals, viewer-centered perspective, and an ability to interact with the virtual world.

Computer-generated stereo visuals allow the user to see the computer-generated world in three dimensions (3D), which is how most (but not all) people see the real world. Each eye sees the world from a slightly different position, allowing us to perceive depth. As with the viewing of stereo photographs or the watching of a 3D movie from the 1950s or 1980s, the trick is to give each eye its own view of the material.

Viewer-centered perspective allows the user to move his body or turn his head and the see the appropriate view of the virtual world from this new position. Combined with stereo visuals, this allows the user to not only see a 3D object in the virtual world but to walk around it or look under it by moving in exactly the same way as a person would move around a real 3D object. In a 3D movie or photograph, the viewing position is static—the viewer sees only what the camera saw. With viewer-centered perspective, the viewer is the camera and always has the correct view of the scene. For this to work, the computer generating the visuals needs to know where the viewer's two eyes are.

There are several different ways to do stereo visuals and head tracking, which lead to different virtual reality display hardware. With a head-mounted display (HMD), the user wears a headset, which isolates her from the real world, with a small cathode ray tube (CRT) or liquid crystal display (LCD) devoted to each eye. This allows the user to turn and tilt her head in any direction and still see the virtual world. A tracker attached to the HMD tells the computer the position and orientation of the user's head. With that information, the computer can determine where the user's eyes are and then draw the graphics appropriately.

A *fish tank* virtual reality system makes use of a computer monitor and a special pair of tracked LCD shutter glasses. The computer monitor displays an image for the user's left eye, at the same time telling the glasses to block out the user's right eye. The computer then does the reverse, showing an image for the right eye while telling the glasses to block out the left eye. By doing this quickly, the user can see objects floating in front of the monitor. The LCD shutter glasses are lighter than a HMD and don't isolate the user from the real world. A tracker attached to the LCD shutter glasses gives the position and orientation of the user's head.

This same technique can be used on a larger scale to create a single, large-drafting-table-size display, such as the ImmersaDesk®. With a larger back-projected display, several people can stand in front of the display at the same time and see the virtual world in stereo, but only one person is head tracked.

Moving from a single large screen to several large screens in a system like the CAVE® allows the user to physically walk around virtual objects. A CAVE typically has three 10-ft^2 walls and a 10-ft^2 floor, although some

Figure 1: The CAVE and ImmersaDesk. (Left) A person in the CAVE wearing tracked shutter glasses to see the virtual world in stereo and carrying the wand. (Right) A user sitting in front of the ImmersaDesk wearing the same tracked glasses and carrying the same wand as in the CAVE. The CAVE and ImmersaDesk users can interact with the same virtual worlds from different perspectives.

CAVEs have four walls, a ceiling, and a floor to completely surround the viewers. This larger space allows five people to comfortably view the virtual world together, although again only one person is head tracked (Figure 1).

Another approach is to take multiple screens and, instead of wrapping them around the user, use them to give the viewer a higher resolution wall made up of several screens. A single screen, whether in a fish-tank virtual reality setup or a CAVE, typically has a resolution of 1280 pixels by 1024 pixels. By combining several screens together, much higher resolutions are possible.

There are many different ways of interacting with the virtual world and many different devices to allow that interaction. The user may want to navigate a large space with a joystick or use a set of buttons to change the properties of the virtual world. Just as the user's head is tracked, other parts of the user's body can be tracked, so the user's body can itself be the interface. It's typical to track the user's hand, or the controller the user is holding, to allow the computer to see where the user is pointing.

Although visuals are the most obvious element of virtual reality, audio is also important, to give the users additional feedback. Haptics, the feeling of touch, is also important in certain virtual reality applications. Often a lack of feedback to one sense is compensated for by feedback to another sense. For example, if you don't have haptic feedback, you may get visual or audible feedback.

In order to keep up the illusion, the imagery of the virtual world must be drawn at a rate of at least 15 frames per second per eye. Otherwise, the world will seem to stutter. In a movie theatre, we watch films composed of still images moving at 24 frames per second and see smooth motion; it's the same in virtual reality. This is why virtual reality requires very powerful computers and graphics cards.

Throughout the 1990s, this required very expensive computers, but now it is possible to do single-screen virtual reality using high-end personal computers. There is also current research going on in autostereoscopic displays, where the user will not need to use special glasses to see computer-generated stereo imagery. For a more thorough discussion of virtual reality, see Sherman and Craig (2002).

COLLABORATIVE VIRTUAL REALITY

In the 1990s, more and more groups around the world gained access to virtual reality equipment, making collaborative virtual reality possible. Again, there are several definitions of collaborative virtual reality—every day many people play collaborative or competitive games on the Internet, which can be considered collaborative virtual reality, and sometimes share environments with hundreds of other players. Since the 1970s, text-based multiuser virtual worlds such as MUDS and MOOS have been popular, evolving from their origins as collaborative adventure games and allowing people to communicate and interact over very-low-bandwidth connections. In the mid-1990s, with advances in both computing power and network speeds, users could explore 3D worlds over the Internet through VRML (virtual reality modeling language) browsers.

For our purposes, collaborative virtual reality requires connecting up the devices described in the previous section, allowing people in several places to share a 3D environment. Some research groups focus on supporting existing low-bandwidth Internet infrastructures or massive connectivity involving thousands of participants at the same time, as in military simulations or Internet-based computer games (Singhal & Zyda, 1999) Our focus on the use of virtual reality for manufacturing, for scientific purposes, and for information visualization has a different set of requirements. We are building systems for small working groups, typically no more than seven collaborators at a time but with large data distribution requirements, to share high-fidelity audio and video communications and large engineering and scientific data stores over high-speed national and international networks.

We want to provide high-quality interaction between small groups of participants involved in design, training, education, scientific visualization, or computational

steering. The ultimate goal is not to reproduce a face-to-face meeting in every detail, but to provide the next-generation interface for collaborators, worldwide, to work together in a virtual environment that is seamlessly enhanced by computation and access to large databases. Although the goal of audio and video teleconferencing is to allow distributed participants to interact as though they are in the same physical location, collaborative virtual reality allows them to interact as though they are the same immersive virtual environment. This way they can interact with each other as well as the objects in their shared environment.

This shared environment may be for designing a new car, visualizating climatological data, or visiting other 3D space that either does not exist physically or cannot be physically accessed. The participants are not talking about a thunderstorm, they are standing inside one; they are not looking at a scale model of a new car design, they are standing inside the full-size engine block. We believe that by transmitting gestures as well as audio and video between collaborators, these shared virtual environments give their users a greater sense of presence in the shared space than do other collaborative mediums. By encouraging collaboration and conversation within the data, these environments may become the preferred place to work and interact even when traditional face-to-face meetings are possible. However, collaborative virtual reality is not going to replace e-mail, phone calls, or existing teleconferencing systems. They each have their strengths and uses. Just as word processing documents, spreadsheets, and white boards shared across the Internet put discussions in their appropriate contexts, so does sharing a virtual space as well as the 3D design being considered or the simulation being visualized. A more thorough discussion can be found elsewhere (Leigh, Johnson, Brown, Sandin, & DeFanti, 1999)

For example, General Motors uses collaborative virtual reality to allow design and manufacturing teams based in several sites around the world to import 3D computer-aided design (CAD) models into the CAVE for quick visual inspection and design reviews at 1:1 scale. The goal is to allow designers to both synchronously and asynchronously access a design that persists and evolves over time from locations scattered around the world rather than forcing collaborators to meet physically at the 1:1 scale clay model. A typical working scenario involves a designer making modifications on a workstation in a 3D modeling package and having those changes propagate automatically to the networked virtual environment, allowing all collaborating participants to see the changes simultaneously. They are then able to critique the design and suggest changes to the designer who can do so immediately at the CAD work station.

The Virtual Reality in Medicine Laboratory at the University of Illinois at Chicago uses collaborative virtual reality to allow a remotely located physician to teach medical students about the 3D structure and function of the inner ear. In this environment, the students and instructor may point at and rotate the ear to view it from various perspectives. They may also strip away the surrounding outer ear and temporal bone to more clearly view the inner anatomy. Audio from the voice conference is used to modify the flapping of the eardrum to illustrate its function. This application is effective because it leverages the stereoscopic capabilities of virtual reality to disambiguate the spatial layout of the various structures in the inner ear—something difficult to do on standard flat images in medical textbooks.

College undergraduates at Central Missouri State University and other universities use Virtual Harlem, a virtual reality reconstruction of Harlem, New York, during the 1920s, in their English classes. Virtual Harlem was designed to immerse students of the Harlem Renaissance directly in the historical context of the literature of that period to reinforce active learning. The goal is to develop rich, interactive, and narrative learning experiences to augment classroom activities for students in the humanities. Collaborative virtual reality allows classes at different universities to meet and share their views within its space, as well as allowing remote expert tour guides to take classes through Virtual Harlem and discuss important issues that the space brings up. This is discussed further elsewhere (Sosnoski & Carter, 2001).

Some virtual environments will only exist while people are inside it; others will be maintained by a computer simulation that is constantly left running. This space exists and evolves over time. Users enter the space to check on the state of the simulated world, discuss the current situation with other collaborators in the space, make adjustments to the simulation, or leave messages for collaborators who are currently asleep on the far side of the planet. For example, in a computational steering application, a supercomputer may be running a large simulation that takes several days to complete. At regular intervals, the supercomputer produces a 3D snapshot of the current data, perhaps a visualization of cosmic strings. A scientist can then step into a CAVE and look at the 3D data that has been produced to see whether the simulation is progressing correctly or whether it needs be tuned, to focus on particular details rather than wait for the simulation to complete.

Avatars

Presence in the virtual world is typically maintained using an *avatar,* or a computer-generated representation of a person. These avatars may be as simple as a pointer that depicts the position and orientation of the wand in the virtual world. However, having representations of the physical bodies of the collaborators can be helpful in aiding conversation and understanding in the virtual space, as you can see where your collaborators are and what they are looking at or pointing at. Tracking the user's head and hand position and orientation allows the computer to draw computer-generated characters representing each of the remote collaborators. These articulated characters move along with the remote user and are able to transmit a reasonable amount of body language, such as pointing at objects and nodding or tilting the head. This style of avatar is useful in task-oriented situations, but do not work as well in negotiations.

Seeing high-quality live video of a person's face can improve negotiations. Video avatars, full-motion full-body videos of users, are realistic looking, which improves recognition of collaborators but require much higher

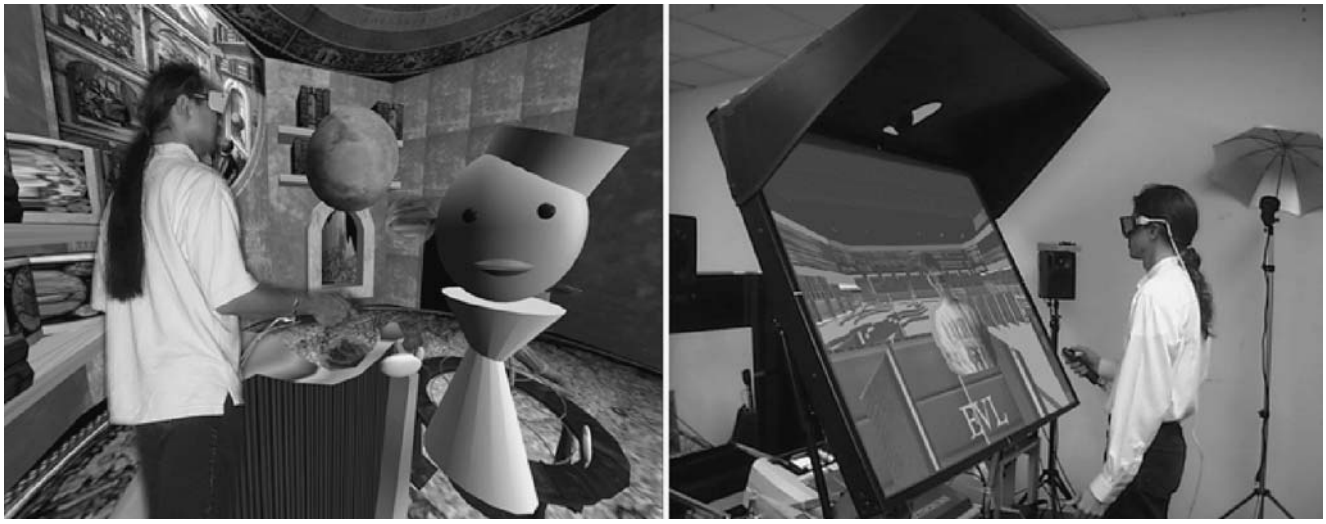


Figure 2: Remote participants in collaborative virtual reality sessions can be seen in the shared virtual space in several ways. (Left) The user is interacting with the articulated computer-generated avatar of a remote participant in the CAVE. (Right) The user is interacting with the live video avatar of a remote participant on the ImmersaDesk2.

network bandwidth and careful setup to achieve high quality. More details on video avatars can be found in Rajan, et al. (2002; see Figure 2). Because the local computer tracks the user's position to correctly draw the virtual world, the computer can send that position and orientation data over the network to remote sites to position an avatar at the same point in the remote virtual environment as the real user occupies in the local virtual environment. The local computer receives updated position information from the trackers many times per second. Because we want to get that information to the other collaborators as quickly as possible, we use an unreliable user-datagram protocol (UDP) connection, rather than a reliable transmission-control protocol (TCP) connection. If one of the packets of position/orientation data is lost, then another will be along shortly. We typically send out avatar position and orientation information 15 times per second. Because each tracker returns three positions (X, Y, and Z) and three rotations (roll, pitch, and yaw), each represented by a 32-bit floating point number, transmitting these values 15 times per second requires only 3-K bits per second per tracker. Typically, there will be at least two trackers in use, one for the head and one for the hand, bringing the bandwidth needs up to 6-K bits per second. On the remote side, the position and orientation of the avatar is usually interpolated from the network data to smooth out the motions of the avatar.

Avatars are also useful in alerting other users to a person's next actions. For example, the declaration "I'm going to move this chair" combined with the visual cue of your avatar standing next to a chair and pointing at it alerts other users to the fact that you are about to grab it. As in real life, we know that if another collaborator tries to grab the chair at the same time, it will be awkward.

Different avatar forms are useful in different virtual worlds. In certain situations, such as pointing at an object on the ImmersaDesk, pointers can be better than full avatar bodies, because, just as in real life, a remote user's avatar body may block your view if you are standing close together. For small objects, short pointers work well, but in larger spaces it helps to have a long beam, allowing users to point accurately at objects 100 feet away.

Avatars with highly stylized bodies are easier to differentiate within the environment but they may not be appropriate for all types of users. First-time users tend to laugh the first time they meet articulated avatars, who appear to be living cartoon characters, but once the characters begin interacting with them, they quickly adapt and have no trouble treating the characters as living persons. Other users desire a more "serious" representation of themselves in the virtual world. Using photographs to generate avatar heads that look like their actual users is a way to help bridge recognition between the virtual world and the real world. Name tags can also be useful in identifying avatars, much as they are useful in real life (see Figure 3).

People feel very uncomfortable if they accidentally walk through another person in the shared virtual world and apologize profusely. In general, people tend to maintain an appropriate distance from other users and try to avoid violating their personal space. Preventing collisions by not allowing one user to move through another can help maintain social comfort. However, this collision detection can be a hindrance if several people are trying to maneuver down a narrow hallway in the virtual world. In certain situations, it is good to have real-world constraints, such as gravity and collision detection, and at other times it is good to be able to turn reality off and be able to do more than you could in real space.

Almost everyone's initial steps into a collaborative virtual world require a tour guide. Initially, this tour guide was a person who would stand next to the new user, show him the controls of the virtual reality hardware, and point out the features of the virtual space. Now, we know that it is useful to have this tour guide give the initial tour remotely as an avatar. This draws the user into the shared virtual space and immediately starts up a dialogue

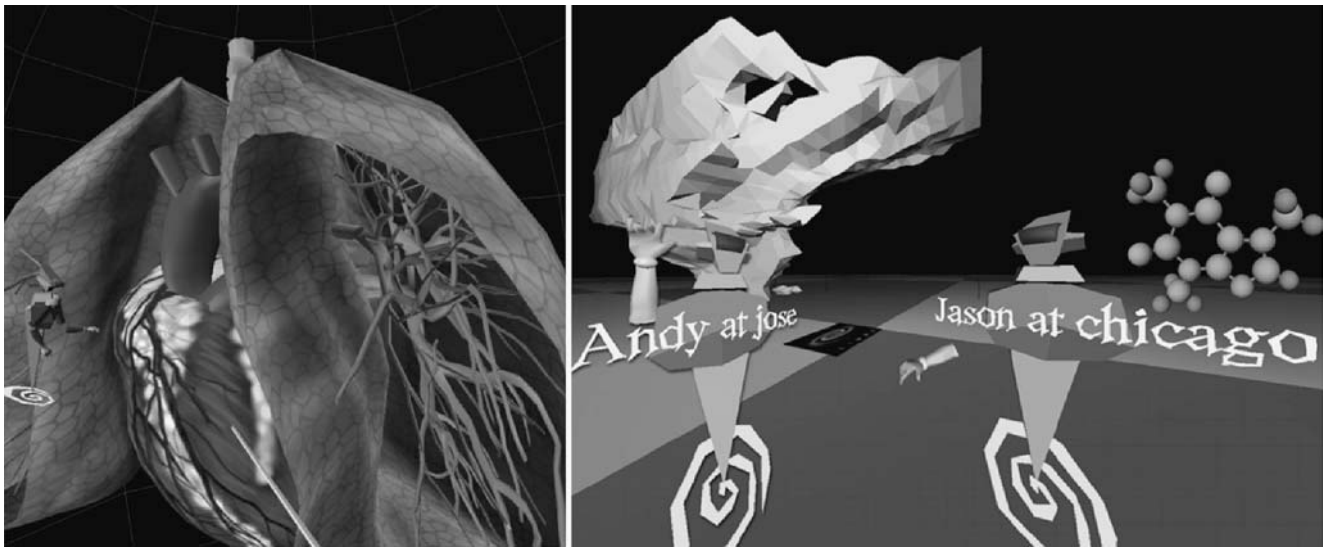


Figure 3: Remote participants sharing collaborative virtual worlds. (Left) Users explore the structure of the human heart at much-larger-than-human scale. (Right) The name tags that help to identify the generic avatar bodies are obvious.

between the new user and one of the users already in the space. This dialogue often deals with how to navigate around and interact with the shared environment, with the remote tour guide showing the new user around their shared space. This way, the new user's attention stays focused on the virtual, rather than the physical, world.

For international collaborations when English is used as the default language, foreign speakers whose first language is not English may find it difficult to converse naturally, and hence these participants tend to be less vocal. Although the tracker is able to transmit gross gestures, it is harder to spot more subtle gestures. What is normally considered a clear nod in the real world usually amounts to a suggestion of a nod in the virtual world. Cultural differences also impact the degree to which a participant gestures. For example, Americans tend to gesture considerably while speaking, whereas the Japanese tend to gesture very little. In situations where collaborators are from different cultures, it may be useful to include video to help mediate discussions so that the faces of the participants can be clearly seen.

Audio and Video

Knowing where your collaborators are and what they are looking at is important, but to really be able to collaborate, a stable high-quality audio connection is needed. If it is difficult to hear the other participants or if there is much delay (latency) in the audio, then the collaboration quickly breaks down. Unlike prerecorded audio over the Internet, where the local computer can buffer the audio and video for smooth playback, here the audio is live, so a delay of more than 200 ms can hurt the collaboration, and high variability (jitter) in the delay will hurt it even more, in much the same way that a bad connection on an international phone call can make a conversation impossible. When multiple people are in the space it can be difficult to tell who is speaking if the audio does not come from a particular point. Adding mouth movements, even crude lip synching, to the avatars gives cues. Directional audio, where the audio appears to come from a specific location in the space, is very helpful, as the voice appears to come from the speaker's avatar. Audio that is quieter with distance can be helpful if there are multiple groups at different places in the virtual space performing different tasks. By sending the participants audio over the network, we are able to alter the volume based on that user's position in the space. Using 8-Khz audio, which gives reasonable quality for conversations, the bandwidth required is 64-K bits per participating site. To reduce the number of encumbrances on the user, we prefer to mount high-quality ambient microphones on the virtual reality hardware itself, though these mikes can have problems in noisy environments.

Sending video information requires much more bandwidth, but it does help in settings where negotiation and recognition are important. Mounting a camera (or several cameras) on the virtual reality hardware aimed at the primary user allows the computer to capture video of the participant while in the space. Sending 15 entire video frames of 720×486 RGBalpha data per second would require 160 Mbps, but much of the information that the camera captures is not relevant. We may be interested in the live video participants face only for mapping onto an avatar at the other end. By reducing the area of interest but maintaining 15 frames per second, we can reduce the bandwidth to roughly 15 Mbps. We can reduce this further through the use of hardware and software compression tools, but the additional latency incurred through compression and decompression can become distracting to the user.

Other Types of Data

Computers running the virtual reality display devices involved in a collaboration need to ensure that the virtual world remains consistent for all users. Application data about the state of the virtual environment is also

transmitted between the various sites. Unlike the avatar data, this application data is typically sent via reliable TCP. In a shared architectural space, this may be the position of each wall and chair. In a scientific visualization, it might be which vectors are turned on.

Typically, all graphics are generated locally at each collaborating site. A new user entering the collaboration may not have all the models being visualized, or she may not have the current versions of the models, so these models will need to be transferred to the local site before she can join the collaboration. These large data files will be sent via reliable TCP.

Haptic devices rely on response times that are too short to allow sending haptic data over networks. Decisions about whether users should feel pressure from the virtual world have to be made locally.

All these types of data moving between various collaborators raises the issue of who has the "correct" version of the world. The simplest solution is to use a central server that maintains the correct state. Each client sends his information to the central server, which broadcasts it to all other clients. If this central server is always left on, then it can also be responsible for maintaining the persistent environment. For larger collaborations, it may be desirable to have layers of servers, to improve local communication. For example, all collaborators in North America would connect to one server, all in Europe to another, and all in Asia to a third. Then these three servers would talk to the main server.

The speed of light can become a limiting factor in these international collaborations. If data is sent around the world via a satellite in geostationary orbit, then the time taken for the data to reach the satellite and come back down can be a significant amount of the total delay (as much as 1 s). This delay is most commonly seen in news reports via satellite from isolated areas of the world. Because of this, whenever possible, we prefer to use fiber-optic connections over the ocean floor for our international collaborations.

Depending on the number of users and the types of data being transmitted between them, the total bandwidth required can be quite large. The various types of data have different requirements. Audio, video, and tracker data are streamed between sites needing low-latency, large-model files are sent occasionally, and at regular intervals small amounts of data are sent to keep the virtual environments synchronized. It is important that these different types of data do not interfere with each other—the tracker data shouldn't be delayed by a large model being downloaded. The current Internet has no quality of service (QoS), the ability to provide guarantees of bandwidth and latency for transfer of information over networks. This means that we need to overprovision the networks to make sure we have excess bandwidth. Only universities, research labs, and large corporations tend to have this necessary bandwidth to spare. Once QoS is available to all, users will be able to customize networks to suit their needs. However, some researchers believe that future generations of the Internet will have so much excess bandwidth that QoS will not be necessary. Regardless of what may happen in the future, collaborative virtual reality applications will always need to adapt to whatever networks they operate over.

SYNCHRONOUS AND ASYNCHRONOUS WORK

When the collaborators are within the same city or within the same continent, it is straightforward to hold synchronous sessions in shared space since time-zone differences vary by only a few hours. When the collaborators are spread across the planet, this becomes much harder. In order to synchronously communicate, one or more groups must either stay up late or get up early. Asynchronous collaboration, where the collaborators work in the same virtual space at different times, allows collaborators to work in shared space during their normal work hours. It can be used to enhance productivity by allowing a collaborator to hand off his work at the end of his day to someone who is just starting her day. There are, however, many cultural and language issues that must be dealt with if these spaces are to be effective supporters of collaboration.

E-mail is a successful tool supporting asynchronous work. However, in international collaborations, there is typically a 1-day turnaround time to get responses, so collaborators can easily waste days clarifying the work to be done and making instructions clear. In a virtual environment, this is even more difficult, as it is hard to use text, speech, or even 2D images alone to describe work to be done or discoveries that have been made in a dynamic 3D environment. It is important that the messages between the distributed team members be clear, to reduce misunderstandings. In a virtual environment, it is important to be able to put these messages in their appropriate context—the virtual world itself.

One advantage of doing design or scientific visualization in an immersive environment is the ability to have geographically distributed participants sharing space with each other and the objects under discussion. This allows the participants to point at specific objects in the scene or set the parameters of the simulation to specific values to clarify what they are saying. It gives the users a common context for their discussions. Especially in international collaborations, where the language barrier can be a large hurdle, being able to gesture relative to the environment (pointing at the red box, turning your head to look at the green sphere) helps to clarify the discussion. The ability to hand off work quickly and accurately is also of great importance. A user stepping into the virtual environment needs to know what work has been done since she was last there, and what new work may need to be done.

When asynchronously working in a virtual space, e-mail and telephone conversations are not enough to describe changes that need to be made or observations that need to be verified. What is needed is a way to allow users to record themselves as avatars talking and gesturing within the virtual space. Then, users in the recorded message would appear just as they would if they were in a synchronous collaborative session. These recordings preserve important head and hand gestures, help clarify the process that went into creating the artifacts in the virtual world, and help support orientation in the space.

These annotations can help explain changes that were made or that need to be made in a collaborative design environment, specify where interesting phenomena were

discovered in scientific visualization environments, and give researchers an easy way of creating metadata recordings from within the virtual space. Most computational scientists agree that a crucial part of the knowledge crystallization process includes the creation of snapshots and annotations to track the progress of the exploration and to record discoveries. In desktop environments, these annotations are typically entered in text windows. However, in an immersive environment, this common mode of data entry is problematic as well as limiting. Current virtual reality displays lack the resolution to display text clearly in a virtual window. Recording and replaying audio messages has been used to try and circumvent these problems. Adding in the avatar of the person recording the message, complete with their gestures, allows those audio messages to be put in the proper context. More details can be found in Imai, et al. (2000)

We have used these annotations in Virtual Harlem to create prerecorded tour guide avatars for visitors. If there is no live expert tour guide available, the virtual world can still introduce itself and take a user on a tour, giving information at various points of interest. The students in the classes visiting Virtual Harlem can also leave annotations in the system—making comments or posing questions to which other users can respond within the space.

Because all state changes to the collaborative virtual world come through the virtual reality application's networking layer, we can store the time-stamped sequence of changes and play back the virtual experience as an immersive movie and experience it from within the virtual environment, watching the action unfold around us. This allows users to record their entire virtual reality experience for playback and analysis. A person playing back the recording will see everything he would have seen if he were present in the immersive space at the time the recording was made. With a video recording of the collaboration, the user is limited to seeing the action from the position of the camera. Here, the user can walk (fly) through the space while the collaboration is underway, watching the action from whatever location seems most interesting. These recordings could also be shared collaboratively during playback, allowing geographically distributed participants to collaboratively watch the recording of a previous collaborative session. For scientists, perhaps the single most compelling reason for recording and archiving collaborative virtual reality sessions is the fact that scientists need to create permanent documentation of their discoveries. Results need to be reproducible and reviewable.

HETEROGENEOUS VIEWS AND ABILITIES

In many existing collaborative virtual reality applications, participants typically all view and modify the same representation of the data they are viewing, but some collaborative virtual worlds allow, and in fact *emphasize,* giving each user a different view of the shared space. These heterogeneous views allow us to leverage the capabilities of a shared virtual space to allow each user to customize his view to his needs.

Different users viewing a multidimensional scientific visualization may be able to partition the dimensions across the various users to break the problem into smaller pieces. Different users may see the same space at different scales. For example, in an architectural space, some users may walk through the space life-size while others see it in miniature, to make it easier to reposition the components of the space. Different users may also have different levels of security access to the data, so some users can see more than others. Individuals who are trying to solve a common problem gather (in workshops, for example) in the hopes that their combined experience and expertise will contribute new perspectives and solutions to the problem. Users may also have heterogeneous abilities in the space, based on their heterogeneous views.

For example, students at Abraham Lincoln Elementary School use collaborative virtual reality with heterogeneous views to learn about the shape of Earth. Two children collaborate in exploring a small, spherical asteroid. One child, acting as an astronaut, explores the surface of the asteroid while the other child, acting as mission control, guides the astronaut from an orbital (spherical) view. Virtual reality helps situate the astronaut on the surface of the asteroid, where she can experience circling the globe and coming back to the same place, not falling off the "bottom," and seeing objects appear over the horizon top first. Virtual reality gives mission control an obviously spherical world to monitor. The two children share the same virtual environment but see it in different ways. They must integrate these different views to complete their mission, and through integrating these views they learn to map between these two different views of the same object. More details on this work can be found in Johnson, Moher, Ohlsson, and Gillingham (1999).

What we have seen so far reaffirms the fact that previous computer-supported cooperative work findings are applicable to collaborative virtual environments. This is discussed further in Churchill, Snowdon, and Munro (2001), who note the following.

- There is a need for individual pointers, to allow collaborators to point at shared data items. However, these pointers can become a source of distraction, and users should have the ability to toggle them on/off.
- It is useful to have some cue as to which region of space a user is manipulating.
- Even in a fully shared environment, participants found the need to work with localized views.
- There is a frequent transition between parallel/independent and coordinated activities.
- The user-interface should be considered part of the visualization, so that collaborators can gain greater awareness of their collective actions as they manipulate the visualization.

The collaborators may also be using heterogeneous devices. Some users, such as those in a CAVE, will have a wider field of view and be better able to see an overview of the environment. Those with higher resolution displays will be better at seeing details. Those with fish-tank systems will have better access to keyboards and other more

traditional interface tools and applications. Collaborators may also augment their virtual reality displays with handheld computers: Laptops, hand helds, and tablets provide additional simultaneous display of additional information.

THE FUTURE

In the 1990s, large, specialized graphics computers were required to drive virtual reality displays. Today, commodity personal computers with high-end graphics cards do very well. The future will see faster PCs with more powerful graphics cards, allowing more people to join in collaborative virtual environments. Currently, for under than $10,000, a person can build a passive stereo projection-based virtual reality system driven by a PC that allows the user to collaborate with others (www.geowall.org). If Moore's law continues to hold true, every 18–24 months these PCs will double in capability while their cost remains constant.

The capacity of high-speed networks is growing even faster than is stated in Moore's law. Currently, network capacity is doubling every 8 months. Such technologies such as dense wave division multiplexing allow multiple light waves to share a single fiber-optic cable. Each of these light waves can provide between 1 and 10 Gbits/s of bandwidth, and each fiber is able to hold hundreds of light waves. Two challenges are emerging from this. The first is to find ways to bring this extreme amount of bandwidth to end users' desktops—what is commonly known as the "last-mile problem." That is, there is an explosive growth of network capacity at the core of the network but no way to extend that capacity to end users. The second challenge comes from a realization that there is more network capacity available than a single user's desktop computer can possibly generate or absorb. This means that future applications of high-capacity networks will consist of clusters of computers communicating with other clusters of computers. The extremely-high-capacity networks that connect these clusters of computers will become the new system bus. The computer will no longer be thought of as the box on your desktop. The desktop computer will simply be the display peripheral for the larger computer that is dispersed around the world, interconnected by high-capacity networks.

For collaborative virtual reality, this will translate into ever more realistic visual experiences, where virtual worlds are populated by synthetic participants who are indistinguishable from human participants. If the virtual experience is compelling enough, some users may prefer to forego physical travel altogether and travel only within cyberspace. For example, they may prefer to take virtual vacations. For people with disabilities that make difficult the rigors of physical travel, this can be a very liberating experience. But even for able-bodied travelers, virtual travel could mean the end of waiting in security lines at airports and sitting in airplanes for hours.

The challenge then will be in developing more compelling experiences or ways for participants to create their own experiences. This is likely to be the future of the video game and movie industry.

CONCLUSION

Our goal isn't simply to make working in collaborative virtual reality environments possible. It's to make it convenient: convenient access to virtual reality hardware, convenient access to virtual worlds, convenient synchronous or asynchronous access to collaborators, and convenient methods for sharing data. As the technology becomes cheaper and more available, more and more people in different domains will work in collaborative virtual reality environments, and more applications will be created and evaluated. Their success will be based not on how they recreate reality but on how they leverage virtual reality to make the collaboration better than being there in person.

GLOSSARY

Asynchronous Collaboration Two or more people cooperating in a task at different moments in time; communication not in real time but through passed messages and shared artifacts.

Avatar A computer-generated representation of a participant in a shared virtual reality environment.

CAVE CAVE automatic virtual environment; a projection-based virtual reality display environment created at the Electronic Visualization Laboratory consisting of three walls and a floor, a tracking system, an audio system, and an interface controller.

Jitter Variability in latency.

Latency The time it takes for a piece of data to get from one site to another.

QoS Quality of service; the ability to provide guarantees on bandwidth and latency for the transfer of information over networks; important when multiple heterogeneous real-time data streams (audio, video, avatar data, control data, etc.) are moving between clients in a collaborative virtual reality environment.

Synchronous Collaboration Two or more people cooperating in a task at the same time and having real-time feedback on their partners' actions.

TCP Transmission control protocol; a reliable Internet protocol that requires more overhead than does UDP, which makes it suitable for transmitting control information in collaborative virtual reality environments.

UDP User datagram protocol; an unreliable Internet protocol that requires less overhead than does TCP, making it suitable for transmitting avatar position information in collaborative virtual reality environments.

CROSS REFERENCES

See *Distance Learning (Virtual Learning); GroupWare; Interactive Multimedia on the Web; Multimedia; Virtual Enterprises; Virtual Teams.*

REFERENCES

Churchill, E., Snowdon, D., & Munro, A., Eds. (2001). *Collaborative virtual environments: Digital places and spaces for interaction*. New York: Springer Verlag.

Imai, T., Qui, Z., Behara, S., Tachi, S., Aoyama, T., Johnson, A., & Leigh, J. (2000). Overcoming time-zone

differences and time management problem with teleimmersion. In *Proceedings of INET 2000: The Internet global summit, Yokohama, Japan, July 18–21, 2000*. Reston, VA: The Internet Society [CD-ROM].

Johnson, A., Moher, T., Ohlsson, S., & Gillingham, M. (1999). The round Earth project: Collaborative VR for conceptual learning. *IEEE Computer Graphics and Applications, 19.6,* 60–69.

Leigh, J., Johnson, A., Brown, M., Sandin, D., & DeFanti, T. (1999). Visualization in teleimmersive environments. *IEEE Computer,* December 1999, 66–73.

Rajan, V., Subramanian, S., Keenan, D., Johnson, A., Sandin, D., & DeFanti, T. (2002). A realistic video avatar system for networked virtual environments. In *Proceedings of immersive projection technology symposia 2002, Orlando, FL, March 24–25*. Ames, IA: Iowa State Universtiy [CD-ROM].

Sherman, W., & Craig, A. (2002). *Understanding virtual reality: Interface, application, and design*. San Francisco: Kaufmann.

Singhal, S., & Zyda, M. (1999). *Networked virtual environments: Design and implementation*. Reading, MA: Addison–Wesley.

Sosnoski, J., & Carter, B., Eds. (2001). *Works and days* 37/38 (special issue on the Virtual Harlem Project). *19 (1–2)*, 1–241.

Virtual Teams

Jamie S. Switzer, *Colorado State University*

INTRODUCTION

> Cyberspace. A consensual hallucination experienced daily by billions of legitimate operators, in every nation, by children being taught mathematical concepts. . . . A graphic representation of data abstracted from the banks of every computer in the human system. Unthinkable complexity. Lines of light ranged in the nonspace of the mind, clusters and constellations of data. Like city lights, receding . . . (Gibson, 1984, p. 51)

William Gibson coined the term "cyberspace" in his popular 1984 novel *Neuromancer*. While perhaps not as dramatically as Gibson's vivid portrayal, the modern world now lives and works in cyberspace. Geography, borders, and time zones are rapidly becoming irrelevant in the way today's business is conducted. The role of technology, particularly the Internet, is profoundly affecting the economy, business, and social fabric of interacting and communicating.

For many reasons, including corporate mergers, globalization, the need to respond rapidly to changing markets and customer demands, increasing sophistication of technology, travel costs, and the trend toward flexibility and mobility in the workforce, organizations must change from the old ways of doing business to new ways. Part of this operational shift is for organizations to have their people perform functional tasks by working in virtual teams. No longer is colocation necessary for group interaction and knowledge sharing. According to researchers Lipnack and Stamps (2000), virtual teamwork is one of the answers to the modern problems of 21st century organizations.

The word "virtual" means "existing or resulting in effect or essence though not in actual fact, form, or name" (*Webster's,* 1984). The concept today has been extended to suggest the use of telecommunications and computing technologies. The Internet in particular has made the use of virtual teams much more effective and desirable. The concept of the virtual organization, while still evolving, relies on the movement of information across cyberspace and new ways of managing people working globally in virtual teams.

VIRTUAL TEAMS

Definition of a Virtual Team

A virtual team is a "group of people who work interdependently with a shared purpose across space, time, and organization boundaries using technology" (Lipnack & Stamps, 2000, p. 18). Virtual teams are geographically dispersed and culturally diverse, often do not have constant membership, and are completely dependent upon technology, particularly the Internet. This is directly contrary to the traditional notion of what makes an effective team. But virtual teams address the needs of the new work environment. The office is where the worker is, not the other way around. Organizations utilizing virtual teams can maximize resources and hire the best people for the job regardless of where they live.

Duarte and Snyder (1999) have identified seven basic types of virtual teams: networked, parallel, virtual project/product development, work/production, service, virtual management, and action. People who collaborate across time, distance, and organizational boundaries to achieve a common goal are part of a networked virtual team. Typically team membership is always changing, with individuals joining and leaving the team as their particular expertise dictates.

Parallel virtual teams also work across time, distance, and organizational boundaries to complete specific assignments, tasks, or functions assigned to the team by the organization. Differing from a networked team in that the parallel virtual team has a consistent membership, team members usually are working together on a short-term basis to achieve a specific objective. Virtual parallel teams are effective when the task involves multinational organizations where a global perspective is needed.

Similar to networked and parallel teams, virtual project or product development teams are created to produce a specific, measurable result, such as a new product. These types of virtual teams work for a designated period of time,

usually of longer duration than a parallel virtual team. Team membership, while it may be fluid, is usually more clearly defined than in a networked team.

Virtual teams that exist to perform one specific function are called work or production teams. With a clearly defined membership operating across time and distance boundaries, virtual work teams carry out regularly scheduled tasks and responsibilities. Team members rarely meet face-to-face, communicating instead via technology tools such as the Internet.

Service teams that provide technical support for organizations are becoming virtual. Each team performs its function during normal operating hours for the team's specific geographic location. When the workday is done for one virtual service team, the task is then passed on to the next designated team in a different location and time zone.

Virtual management teams work across time and distance but do not cross organizational boundaries. Members of these teams are located around the world yet work together on a daily basis. Multinational and global organizations have their management teams communicate using the Internet and other technologies to accomplish their goals.

Action teams are the final type of virtual team. When an immediate response to a specific, one-time event is required, virtual action teams can provide information and communicate with a variety of people to resolve the issue. Action teams by definition must be working in the same time period but do cross distance and organizational boundaries.

Characteristics of a Virtual Team

A virtual team may exist for a short period of time or a much longer duration. It can be composed of individuals solely from within the organization or a mixture of people from several different organizations. The virtual team may work only with one special project, or simply concentrate on routine tasks. Regardless of the nature of the virtual team, they all share common characteristics.

First and foremost, communication in the virtual team is always mediated by some form of technology. Without the ability to communicate effectively, virtual teams would not exist. The development and increasing sophistication of the Internet and other communication technologies has made the use of virtual teams possible.

By definition, virtual teams are separated by physical distance, in different geographic areas and various time zones. Work can continue 24 hours a day, seven days a week, essentially following the sun around the world. The use of the Internet and other communication technologies makes this global collaboration possible.

A virtual team always has a common task or goal to work toward. This goal is the guiding force in the productivity of the team. Virtual teams are result-driven, and all the team members work together to accomplish the ultimate objective.

Creation of a Virtual Team

The creation of a virtual team involves several different steps. Before any work can begin, five tasks must be accomplished at the beginning of the process. The success or failure of the virtual team depends on executing each step effectively.

Identify the Need for a Virtual Team

The first step is to determine if there is a need to create a virtual team. There are many reasons organizations utilize virtual teams. Mergers and acquisitions of new businesses can create circumstances where an organization has many different units scattered throughout the world. A company can also form alliances and partnerships with international entities, creating the need for a virtual team. Globalization and worldwide competition is creating increasingly complex business situations necessitating the establishment of virtual teams.

Changes in the marketplace have also driven the development of virtual teams. There is an increasing amount of specialization, which in turn heightens customer expectations and provides more alternatives. Organizations may turn to the use of virtual teams to stay competitive. Scarce resources could be distributed and shared for a possible cost savings. Outsourcing some functions may also enable an organization to maximize its strengths and concentrate on its core competencies.

Societal changes have had a role in the proliferation of virtual teams. Time compression in business and industry has created an environment where work must continue 24 hours a day, seven days a week. Employees are increasingly demanding flexible schedules; telecommuting is becoming more of an option. Working in a virtual team provides the solution to some of these societal issues.

Getting the job done in the fastest, most efficient way is how organizations survive and prosper. With the use of virtual teams, businesses can procure experts to participate in the process no matter where they are located. If an organization does not have people in its immediate vicinity who are subject matter experts, the need to create a virtual team is urgent and substantial.

Identify the Virtual Team Members

After the need for a virtual team is identified, the makeup of the virtual team must be determined. The expertise needed to successfully complete the task must be ascertained, and the best people for the virtual team identified. The organization has to ensure that all key areas are represented.

A virtual team consists of a core group of heavily involved people who carry the bulk of the task. Then there are additional people who operate as an extension of the core group, who contribute their knowledge and expertise on an as-needed basis. Often in virtual teams one person plays multiple roles and has more than one function within the team.

When identifying members of the virtual team, an organization must also consider the less-obvious qualifications of a person. The team member must be able to work efficiently in a virtual environment. It is critical that the virtual team member be competent in the use of the technology to be a successful addition to the team.

Identify the Virtual Team Leader

The task of being an effective leader in a virtual environment can be daunting. The nature of business today demands more from virtual leaders than traditional leadership roles. To be a successful virtual team leader requires a special set of skills. More than just computers and technology are needed. Virtual team leaders must have the tools, techniques, and strategies that work in a virtual environment.

Organizations that formally identify virtual team leaders often make a long-term commitment to those leaders by providing resources, training, and other types of support. Savvy management selects virtual team leaders for their leadership and facilitation skills. Yet even if they are formally identified, often people are thrust into leading a virtual team with little or no training and/or support from the organization. Many virtual team leaders report their biggest challenge is an increased sense of burden and responsibility.

There can also be an informal identification of a virtual team leader. The organization may not specifically select a virtual team leader, yet one may evolve or emerge. Such an informal process, however, may be detrimental to the success of the virtual team in that valuable time is lost while the members of the virtual team struggle to determine a process for operation.

Because of the diversity of technical and management expertise needed for virtual teams to be successful, shared leadership is often the norm rather than the exception. There may be a rotation of the overall leadership role in a virtual team, with most members of the virtual team taking a leadership role at some point in the process. Virtual teams that deal with complex issues and problems usually have shared leadership regardless of the titles used. They establish a division of leadership. Each virtual team member has a unique set of skills that contributes to the successful completion of the task at hand.

Identify the Task

The next step in the creation of a virtual team is to identify the task the team will be charged with. The virtual team may be assembled to work on one specific problem or product. Virtual teams can also work with routine, ongoing tasks.

The task must be clearly identified, with the processes involved described in detail. The virtual team must determine the flow of the work, which does not have to be sequential, and assign duties to each member of the team. The specific actions that the virtual team members undertake to accomplish the task will determine the desired end result.

The virtual team must carefully plan how it is going to successfully complete the task. Timelines and milestones need to be determined, and everyone must agree on what the results will be. All virtual team members should have a shared understanding of the task, their individual roles, and who is accountable for what. Objectives and goals must be clearly defined and accepted by all members of the virtual team. Methods need to be developed to review the virtual team's progress and results.

Create Procedures and Processes to Achieve the Common Goal

There are numerous challenges of culture, geography, technology, and organization inherent in virtual teams. Successful virtual teams know that each member has a task to perform for which each will be held accountable. Operating procedures and processes need to be developed and implemented to contribute to the seamless functioning of the virtual team. Norms must be established. Each member of the team should be involved in making key decisions.

Virtual teams operate in an adaptive, volatile environment. Placing rigid controls and restrictions on the virtual team will destroy the team's ability to perform at its highest level. The virtual team must create an infrastructure appropriate to the organization, the task, and the team, which allows the virtual team to function at its highest capacity.

The Internet needs to be considered as a vehicle for communication, not just exchanging information (Gundry, 2001). Virtual team members need to be in constant communication to successfully complete the task and attain their common goals. Effective communication can be achieved by utilizing the Internet to its fullest capacity. With such a variety of tools available the Internet can facilitate all virtual team functions.

The team must have a clear, shared vision, mission, and strategy based on the organization's values and principles. The virtual team must define its expectations and its environment so the virtual team members will have a clear understanding of how the task will be completed and in what order. Outlining the scope and responsibilities of each team member, implementing performance measures, and determining methods needed to review the virtual team's progress and output are also necessary to ensure success.

Because virtual team members usually do not share a common physical workspace, an infrastructure must be created that allows the virtual team to develop and maintain a sense of identity. Strong symbols are needed to unite people across time and space. This can be accomplished in many different ways, most notably using the Internet.

A short, informal Internet video conference would allow the virtual team to build rapport. The virtual team could create a symbolic name for the team and design a representative logo. A designated virtual team historian could maintain a Web site with the location, photograph, and personal information of each virtual team member. An Internet-based "electronic water cooler" (Haywood, 1998, p. 36) could provide a forum where virtual team members could have electronic chats, exchange ideas and information, or just keep in touch.

A method for providing timely, succinct, and effective feedback needs to be in place. This can be accomplished by using a variety of Internet technologies. Additionally, rewards and recognition, along with celebration of achievements, keep virtual team members connected and convey a sense of progress and team identity.

Cultural differences can easily create major misunderstandings within a virtual team. For instance, different cultures may be more or less inclined to provide feedback

using e-mail. Reluctance to communicate using e-mail could result in a decrease in efficiency of the virtual team, so it is the responsibility of the virtual team leader to encourage the acceptance of all types of cultural styles. An understanding of the cultural differences among virtual team members can be used to the team's advantage.

Different cultural groups require different styles of management and organization. The team must develop an approach that fits the mix of cultures the virtual team embodies, taking care not to elevate one culture over another. Virtual teams can create ways of working and interacting that not only accommodate, but also optimize cultural differences among the virtual team members.

One of the foundations for the success of a virtual team is to build trust. This requires a conscious and planned effort on the part of the virtual team, because the virtual team members do not have the traditional ways to meet each other in person and develop a relationship. Every task, action, and initiative builds trust.

The virtual team must ensure respect, fairness, and equality of opportunity for all the virtual team members at all times. This can be accomplished by ensuring that each member of the virtual team has a chance to contribute and to receive feedback. Treating all virtual team members with the same consideration, and allowing the virtual team members time to interact with each other on the same level, can also build virtual team trust.

Clear lines of communication are essential for the success of a virtual team. Because virtual teams rely completely on mediated communication, it is critical to determine which technology is right for the team and the task at hand. The virtual team needs to plan for the use of technology given the team's task, the skills and backgrounds of the virtual team members, and the sophistication of the organization.

Different technologies facilitate different types of communication. Regardless of the technology used, communication must flow freely and frequently, with everything completely understood by all the virtual team members. Specific Internet-based communication technologies have to be identified to ensure effective communication.

TECHNOLOGICAL INFRASTRUCTURE NEEDED FOR VIRTUAL TEAMING

The development of the Internet and the increasing sophistication of additional communication technologies have made virtual teaming possible. The very nature of virtual teams creates a challenge to effective communication. Because members of the virtual team are separated by time and distance, communication among team members must be constant and clear. Messages must be acknowledged, understood, and acted upon.

Virtual team members do not have the opportunity to communicate informally in the hallway of a building or in a parking lot. Therefore, it is critical that communication in virtual teams must be highly structured to ensure that the team members are receiving all the information necessary to complete the assigned task. Virtual teams use a variety of technologies to communicate. The Internet, e-mail, video conferencing, bulletin boards, and groupware are just a few of the tools used by team members to keep in constant contact. Software that allows the virtual team to have shared access to company services as well as the knowledge collected and developed by the team also plays a part in successful virtual team communication.

Virtual teams have four different options in which to work collaboratively: same time/same place, same time/different place, different time/same place, and different time/different place. The team members can be in the same geographic locale at the same time, or they can be in different locations at the same time. This type of communication is called "synchronous" communication and includes such Internet technologies as phone calls, instant messaging, video conferences, and chats.

If virtual team members are communicating in an "asynchronous" manner, they are separated by both time and distance. They can be in different geographic locations at different times, or the virtual team members can be colocated in the same place but at different times. Examples of asynchronous Internet technologies include e-mail, voicemail, Web sites, databases, and bulletin boards. Internet-based groupware can be both synchronous and asynchronous depending upon how the virtual team uses the technology.

Synchronous Communication Technologies

While the majority of communication among virtual team members occurs asynchronously, there are instances when synchronous interaction is more effective and efficient. Using synchronous Internet technologies, virtual teams come together at the same time, but not necessarily in the same geographic location. A simple face-to-face meeting held in the organization's conference room, or a cell phone call, are examples of synchronous communication. More sophisticated types of synchronous communication technologies are Internet-based and include electronic chats (also called instant messaging), video conferencing, and Web-based conferencing.

Chats

Communicating synchronously using the Internet by typing text messages between virtual team members is known as chatting or messaging. The interaction occurs in real time. Team members, using a computer and Internet connection, exchange text messages and have an electronic "discussion."

Video Conferencing

Video conferencing via the Internet can be done using a desktop or laptop computer, or using specially designed video conferencing units. Equipped with a camera, microphone, and Internet access, virtual team members can see and hear each other while sitting at their computer. Larger, more sophisticated Internet-based video conferencing technology is usually housed in a specific area of an organization and not at the virtual team member's desk. While video conferencing allows the virtual team to interact visually and aurally, a mechanism for sharing documents must be in place at the time of the video conference. This can be accomplished using a fax machine or electronic white boards.

Web-Based Conferencing

Virtual teams can hold audio conferences using the Internet by using voice over IP (VoIP). VoIP is the transmission of telephone calls over the Internet. There are four ways to communicate with VoIP, using a computer and/or a telephone.

Virtual team members can use computers alone to conduct an audio conference. All that is required is the necessary software, a microphone, speakers, a sound card, and an Internet connection. This form of VoIP allows for the integration of other Internet applications, such as e-mail and application sharing.

The computer-to-telephone method permits calls from a computer to any standard telephone. Calls from a telephone to a computer are also possible with the appropriate hardware and software. VoIP can also be used in a standard telephone-to-telephone exchange. This method does not involve a computer but is usually more cost effective than traditional long-distance calling.

Asynchronous Communication Technologies

Asynchronous communication in virtual teams does not occur in real time. Any interaction between the team members is delayed. Most communication among virtual teams is asynchronous with a variety of different technologies utilized. The majority of tools are Internet-based, because the Internet is the most ubiquitous technology available to virtual teams. All a team member needs is a browser and Internet access to be immediately connected to the rest of the virtual team.

E-mail

E-mail is by far the most common means of communication among virtual teams. Team members can communicate electronically from practically anywhere in the world and share information. Distribution lists can send a message to every single person on the team, or a subset of the virtual team. Filtering capabilities make it easy for virtual team members to organize their e-mail messages. E-mail is also the most cost effective communication technology for virtual teams to utilize.

Virtual teams must establish a set of rules for e-mail communication to be effective. The e-mail addresses of each virtual team member should be provided to every member of the team. The virtual team needs to establish how often members will check for e-mail messages and what is the expected timeframe to respond to an e-mail. The virtual team members determine what type of information will be included in an e-mail, and how detailed that message will be. If attachments to e-mail messages will be routine, the virtual team needs to ensure that everyone has the proper tools to open the attachment.

It is of particular importance that virtual teams provide guidelines for what constitutes acceptable behavior when using e-mail. The nature of the technology makes it very easy to send an e-mail without thinking through the consequences, which could be devastating to the team and its purpose. Additionally, virtual team members should practice good "netiquette," an informal set of rules designed to regulate conduct on the Internet. For example, proper netiquette dictates that e-mail messages should not be typed in all capital letters. On the Internet, that indicates "shouting" and can be offensive to those who receive the e-mail message.

Human beings use more than words to communicate. Subtle changes in body language can contradict the actual words being communicated. Because e-mail is strictly a text-based computer-mediated communication, written words alone could easily be misinterpreted by members of the virtual team. "Emoticons," created using keyboard characters, are often used in e-mail communications to convey emotions or visual cues lacking in the words alone. A smile, for instance, can be added to an e-mail by typing three characters: a colon, a dash, and a close parenthesis.

Web Sites

Maintaining an Internet presence for the virtual team as a whole, and individual Web sites for each member of the team, is a very effective method of asynchronous communication. Personal Web sites posted on the Internet can contain a wealth of content vital to the virtual team's success. Team members can communicate their contact information, such as their phone number, fax number, e-mail address, and availability. Vast amounts of information such as project updates, schedule changes, and status reports can be accessed by the entire virtual team through its members' individual Web sites. Internet bulletin boards can also be a part of the virtual team's Web sites, allowing team members to exchange e-mail and access information.

Intranets

An intranet is similar to the Internet, but accessible only to authorized users, most often within a particular organization. The intranet is protected from outsiders by a firewall, which denies unauthorized people access to the system. Intranets deliver news and information specific to the organization, provide e-mail distribution lists, and offer bulletin boards and electronic forums for information sharing, project planning, and status reports. Intranets are very functional tools for virtual teams to utilize, allowing open lines of communication and data exchange in a relatively secure electronic environment.

Databases

Databases are used for information storage and retrieval. Copious amounts of valuable data can be accessed by the virtual team members from anywhere at any time through the Internet. Databases are another valuable tool for asynchronous communication in a virtual team.

Groupware

Some tools used by virtual teams to communicate and collaborate via the Internet belong to a class of products called "groupware." Groupware is a generic term for specialized computer aids designed for the use of collaborative work groups (Johansen, 1988). Concepts similar to groupware are often called electronic meeting systems, computer supported cooperative work, computer assisted communication, or group decision support systems. The evolution of groupware products has also aided in the development of virtual teams. Without electronic

collaborative tools, working in virtual teams, uncoupled from time and space, would be extremely difficult.

Internet-based groupware allows virtual teams to work together to perform many different types of functions. Groupware products offer a combination of asynchronous and synchronous communication and information exchange tools in one package. Using groupware, virtual teams can send and receive e-mail, access databases, conduct video conferences, brainstorm, chat, have discussions in electronic forums, use bulletin boards and white boards, execute project management, schedule events with a calendaring function, and share documents and applications.

Groupware is becoming more and more sophisticated as the technologies have evolved. Using just a Web browser and a telephone, virtual teams can access a URL at a scheduled time to have real-time audio chats while simultaneously working on shared documents. Virtual teams can conduct meetings in their own virtual world. These artificial environments are created with computer hardware and software. The groupware allows the virtual team members to represent themselves in this virtual world as either an image or icon (in a two-dimensional virtual world) or as an avatar (in a three-dimensional virtual world) so it looks and feels as if the team members are actually sharing the space.

Choosing the Correct Technology

Virtual team members must work with each other to share information, brainstorm ideas, receive feedback, make decisions, get training, and learn new techniques. Different technologies facilitate different types of communication. The virtual team must select the most efficient and effective tool to communicate given the team's task and type, the skills and backgrounds of the virtual team members, and the resources of the organization.

There are other factors that assist in determining which communication technologies the virtual team will utilize. Cost of the hardware and/or software needed for some technologies can be prohibitively expensive. A technology may have a steep learning curve and the amount of training required would not be feasible for the virtual team. Additionally, there may be some team members who are "technophobes" and not comfortable working with any type of technology.

The organization must have the appropriate physical infrastructure. More sophisticated communication technologies such as Internet-based video conferencing need a higher bandwidth to operate effectively. The hardware/software has to be compatible with the already existing systems. Technology can—and will—fail. Virtual teams need a backup plan if one of the primary technologies becomes inoperable; access to technical support is critical. Having a previously agreed upon method for urgent communication in the case of technical failure should be established by the virtual team at the very beginning of the process.

The technology must match the goals of the virtual team and facilitate what the team is trying to accomplish. Regardless of the technology utilized, the communication among virtual teams has to be very focused and organized. Protocols, such as the primary tool used for communication, plus the language, timing, member accessibility, prioritization, and response rate, need to be established for the virtual team to function effectively.

Synchronous communication tends to build relationships in virtual teams more quickly than asynchronous interaction (Haywood, 1998). Positive feedback can be delivered using any type of technology, but negative feedback should be discussed where the team members' faces can be seen, such as in an Internet video conference. Video conferencing also helps the team create a sense of identity because it is a technology with a high degree of social presence, meaning the virtual team members can connect on a much more personal level than using a technology such as e-mail, which has a much lower degree of social presence.

IMPACT OF VIRTUAL TEAMS ON THE ORGANIZATION

The growth of the Internet and other communication technologies, along with a variety of sociological and economic forces, has made virtual teaming a reality. More and more organizations are embracing virtual teams and making them part of everyday operations. There are both advantages and challenges, however, in working with virtual teams.

Advantages

The use of virtual teams provides the organization more flexibility. No longer bound by time and distance, teams can come together as fast as the mouse click that logs them on to the Internet. The organization has more latitude in how, when, and where tasks are performed. Virtual team members are also afforded more flexibility on a personal level. Working virtually, people are not tied to a specific place or timeframe in which to perform the task, allowing the team members to organize their time in the most efficient and effective way for each of them personally.

Because virtual teams can consist of members from all over the world, there is a larger group of people to draw from. The diverse makeup of the virtual team provides for more ideas from a variety of viewpoints. With virtual teams organizations can also utilize expert consultants and bring in team members with specific expertise more cost effectively. There is also a much greater opportunity for information sharing and collaboration among virtual team members using Internet-based communication technologies, along with more occasions for input and feedback from all virtual team members.

There are chances for creating strategic partnerships or alliances that may not have been feasible without employing virtual teams. Organizations may become more competitive. An increase in the product cycle development or reduced time to market is possible when using virtual teams. Utilizing virtual teams can also provide an organization the opportunity to increase productivity. Internet-based communication technologies allow virtual teams to essentially work around the clock from any geographic location. The lack of common distractions that can occur

in a traditional office setting may also increase the virtual team's focus on the task.

Organizations may recognize significant cost savings by making the most of what Internet technologies have to offer. Not having team members physically located at a central facility can save on office space. Virtual team members interact using Internet-based communication technology, eliminating the need to travel for face-to-face meetings. With virtual team members not traveling long distances, as well as not driving to a fixed office location, the use of virtual teams is also very environmentally friendly.

Lastly, the use of virtual teams can be a great advantage in the event of a catastrophe. Natural or human-made disasters can be devastating to organizations that have their entire workforce in one physical site. Virtual teams are spread out around the globe, ensuring that data and other vital functions can continue in the event of a calamity.

Challenges

Virtual teams must address a variety of challenges that can be detrimental to the organization. While being separated geographically can be an advantage for virtual teams, it can be a disadvantage as well. If a face-to-face meeting is required, the logistics of arranging such a gathering can be cumbersome. Even an Internet-based synchronous meeting can be difficult to organize due to time zone differences.

Often virtual team members will feel isolated, lonely, and disconnected from the organization and consequently will not form a productive, cohesive working unit. They may fear their accomplishments will not be recognized by the organization. Conversely, the line between work and home can become blurred for virtual team members. The constant communication among the virtual team can result in information overload. It is possible that virtual team members will lose focus on the task if they are not self-directed personalities.

Miscommunication can happen very easily in virtual teams. Cultural differences can be barriers to effective communication among all members of the team. Not everyone may speak the same language, resulting in misunderstandings that can be damaging to the organization.

Using Internet-based communication technologies can also result in an increase in costs for the organization. Technology is very resource intensive. Initial purchase of the hardware and software, upgrades, maintenance of the infrastructure, training of employees, and the creation of a support team can be an ongoing expense that is detrimental to the organization. New technologies may not be compatible with existing technology, resulting in further cost to the organization. Keeping the Internet-based technology used by the virtual teams secure from outside destructive forces such as hackers requires further outlay of revenue.

If the administrative structure of the organization does not support the concept of virtual teams, the team will fail. Working in a virtual team requires a new way of thinking for both employees and management. Organizations that do not provide the tools necessary for the success of the virtual team will be at a distinct disadvantage.

CONCLUSION

The virtual world is a reality. More and more people are interacting at all levels—economically, socially, and educationally—solely in a virtual environment. Because of the nature of business today, using virtual teams is rapidly becoming an industry standard and organizations must adapt.

Successfully implementing a virtual team demands careful planning and preparation. The need for a virtual team must be well established. The virtual team members and the leader of the team need to be identified. A task for the virtual team and the processes to complete that task have to be clearly articulated. The challenges of culture, geography, technology, and resources must be dealt with accordingly.

To keep the virtual team operating efficiently there need to be very clear and constant lines of communication between all involved. A strong technological infrastructure, primarily Internet-based, is critical for the virtual team to be effective and accomplish its goals. The variety of Internet tools available to virtual teams ensures that the proper technology is used in the facilitation and ultimate completion of the task.

Without the Internet, virtual teams would not exist. The use of synchronous and asynchronous Internet-based communication technologies has allowed people to work and collaborate without regard to time or geography. New and developing Internet technologies are increasing the amount and quality of interactions between virtual team members. The use of the Internet in virtual teams is becoming seamless and practically invisible—the Internet is simply a part of the everyday operation of a virtual team.

GLOSSARY

Avatar A graphical representation of a person interacting in a virtual world.

Bulletin board A groupware product that allows for electronic exchange of information, chatting, and document sharing.

Chat Communicating in real time by typing text messages on a computer; also called instant messaging.

Cyberspace The world that exists on the Internet.

E-mail Messages sent electronically over communications networks.

Emoticon Using keyboard characters to textually convey an emotion.

Firewall Software preventing unauthorized access to a computer system.

Groupware Using computer aids to work collaboratively.

Internet A decentralized networking infrastructure connecting computers around the world.

Intranet The same technologies as the Internet, existing only within an organization behind a firewall.

Listserv A mailing list that transmits e-mail messages available only to people on the list.

Netiquette Appropriate behavior by people using the Internet.

Remote access The ability to connect to a network from a distant location.

URL Universal resource locator, the address of resources on the World Wide Web.

Video conference Use of video and audio to connect in real time in geographically dispersed sites.

Virtual reality Use of computers to create the illusion of a place or time.

Virtual team A group of people working together separated by time and distance.

Virtual world An artificial environment created by a computer giving the illusion of an actual place.

World Wide Web A way of accessing information over the Internet.

CROSS REFERENCES

See *GroupWare; Internet Literacy; Intranets; Online Communities; Virtual Reality on the Internet: Collaborative Virtual Reality.*

REFERENCES

Duarte, D. L., & Snyder, N. T. (1999). *Mastering virtual teams: Strategies, tools, and techniques that succeed*. San Francisco: Jossey-Bass.

Gibson, W. (1984). *Neuromancer*. New York: Ace Books.

Gundry, J. (2001, June). Managing through the Internet. Retrieved April 23, 2002, from http://www.knowab.co.uk/wbwmanage

Haywood, M. (1998). *Managing virtual teams: Practical techniques for high-technology project managers*. Boston: Artech House.

Johansen, R. (1988). *Groupware: Computer support for business teams*. New York: The Free Press.

Lipnack, J., & Stamps, J. (2000). *Virtual teams: Reaching across space, time, and organizations with technology*. New York: Wiley.

Webster's II new riverside university dictionary. (1984). Boston: Houghton Mifflin.

FURTHER READING

http://www.virtualteamsresearch.org
http://www.icasit.org/virtualteams.htm
http://www.marshall.usc.edu/ceo/vt
http://www.virtual-organization.net
http://www.ascusc.org/jcmc/index.html

Visual Basic

Dennis O. Owen, *Purdue University*

INTRODUCTION

The information systems industry has been profoundly impacted by the Internet. The Internet is an expansive wide area network, one that is more far-reaching that anything that could be constructed by a single organization, or even a group of organizations working collaboratively. A plethora of software is available to manage information and communicate on the Internet. In the true spirit of capitalism, this software has become sophisticated, fast, and relatively inexpensive. This combination of powerful, inexpensive data management software and expansive wide area networking is something business and industry must seriously consider as a viable information distribution system.

As Internet technologies are adopted in business information systems, it is necessary to supplement the available software with equally powerful applications that address unique data management needs within a business. Development of these applications requires a tool that facilitates development, testing, and deployment with a minimum investment of time. Visual Basic, by Microsoft, is such a tool. Visual Basic provides a visual user interface to speed application design. The integrated development environment (IDE) provides many powerful debugging tools to speed development and testing. Visual Basic provides a set of commonly used controls that can easily be incorporated into applications. Finally, Visual Basic provides a complete set of utilities for easy packaging and deployment of new applications. The latest release, Visual Basic.NET, supports Microsoft's new .NET initiative. This seamless integration of the application development environment into the Internet's basic information-sharing and communications framework makes Visual Basic an excellent choice for Internet technologies application development. Currently, over three million developers worldwide use Visual Basic to create applications and other software components (Shelly, Cashman, Repede, & Mick, 1999).

VISUAL BASIC—THE LANGUAGE

The Evolution of Visual Basic

Visual Basic traces its roots back some 40 years to the BASIC programming language developed by Dartmouth College professors John Kemeny and Thomas Kurtz. BASIC was originally intended as an instructional language. The 1970s brought a general decrease in computer costs and a corresponding increase in college computer courses. BASIC, recognized for its wide range of features, became a favorite of students. Upon graduation, these students carried BASIC into business and industry.

Business and commercial users of BASIC soon found that the original Dartmouth specification lacked certain desirable features. Among these were text editing, printed output formatting, and file handling. BASIC vendors began to add to the original Dartmouth specifications to address these shortfalls. The addition of these extensions began the evolution toward today's Visual Basic (Hare, 1982). Currently, the most widely used version of Visual Basic is 6.0. Microsoft recently released a new version of Visual Basic, Visual Basic.NET, which incorporates its new .NET Internet technologies. Visual Basic.NET and its relationship to Visual Basic and Internet programming will be covered later. Visual Basic comes in several different editions designed to meet the needs of specific classes of users. The Learning Edition is a student version of Visual Basic that supports the language elements necessary for development of simple Visual Basic programs like those used in college-level course work. Anyone planning to use Visual Basic for commercial or professional applications development should use either the Professional Edition or the Enterprise Edition. These editions support

advanced database management and offer tools to package and distribute applications. (Bradley & Millspaugh, 1998).

Visual Basic as a Structured Language

Visual Basic is a structured programming language. Structured programming, which grew into its own during the 1970s, is a disciplined approach to programming that strives to make programs easier to understand, develop, test, and maintain. Logic in a structured program can be expressed only through the use of restrictive control structures, each of which has a single entry point and a single exit point. These control structures include the following:

Sequential structure,
Conditional (or selection) structure,
Repetitive (or iterative) structure,
Subroutine and function calls, and
Error trapping.

Visual Basic implements these control structures with a variety of statements. The conditional, or selection, structure is implemented with If, If . . . Else, and Select Case statements. Visual Basic also supports five repetitive structures, the pretest While, the pretest Until, the posttest While, the posttest Until, and the counter-controlled (For . . . Next) loop structure.

Structured programming also encourages the use of modularization, dividing large programming tasks into smaller, more manageable segments. These segments, called modules or subroutines, are supported in Visual Basic by procedure calls and function calls. The use of modules also promotes the concept of reusable code, sections of a program that can be accessed several times and from several locations within a program (Harriger, Lisack, Gotwals, & Lutes, 1999).

Error trapping in Visual Basic is accomplished through the Err object. This object, which exists in all Visual Basic programs, contains properties that are updated with specific information about all errors that occur. Specific sections of a module can be designated as error handlers by the On Error GoTo statement. When an error occurs, program control is transferred to the segment of the module specified in the GoTo portion of the On Error statement. A module may handle some errors internally and pass others to the operating system for disposition (Harriger et al., 1999).

Visual Basic as an Application Development Tool

Rapid Application Development

The goal of RAD (rapid application development) is to speed and streamline the design, coding, testing, and deployment of applications. Visual Basic is well suited to this task. Visual Basic provides the programmer with a comprehensive list of controls that are commonly found in Windows applications. These controls can be quickly placed in a Windows application with a simple drag-and-drop action. The controls immediately exhibit the look, feel, and operating characteristics of similar Windows controls. The repetitive and tedious task of recreating each control has been eliminated, allowing the programmer to quickly build the interface and move on to the task of making the various controls react appropriately to user manipulation. Since a control's behavior is consistent across modules and between applications, maintenance and team application development are easier (Shelly et al., 1999).

Rapid Interface Design and Prototyping

The rich control set provided by Visual Basic is an asset in the prototyping and design phases of program development. The ability to quickly design and build a partially functional user interface allows the programmer to give quick feedback to end users about the information used, the information created, and the way users interact with the program. From this, users can determine if the application meets their needs before valuable resources are spent on the complete design and development process. Consistent control behavior across applications allows users to be productive with little or no training. Users can interact with the program while it is still in the early design stages and provide suggestions for modifications while they are relatively easy to implement. With minimal programming, controls can respond to user inputs. Users get a better idea of what the program does and how it "feels." Users can identify problems with the data and the interface early in the development process, while changes are relatively easy to make. Early user involvement in the testing process prevents the investment of development resources in a system based on incorrect processing, incorrect data, or ineffective interfaces. The more efficient use of resources also reduces development time.

Wizards

Visual Basic provides several wizards to speed many of the common tasks related to application development and deployment. Visual Basic includes wizards to build application frameworks, create data forms to display information from a database, and create setup disks to deploy Visual Basic programs (Harrington, Spenik, Brumbaugh, & Diamond, 1997). Wizards automate a task by using a series of dialog boxes to gather information and walk the user through the process (Harriger et al., 1999).

Object-Oriented Programming and Visual Basic

Visual Basic is frequently called an object-oriented programming (OOP) language. A basic understanding of objects is needed to fully understand how Visual Basic fits into the OOP framework (Bradley & Millspaugh, 1998). An object can represent anything, from an automobile to a microware oven to a mathematical equation. All information about the object, including its uses and behaviors, is part of the object. An object is, in a programming environment, a set of data and any programming needed to manipulate those data. Combining data and programming into a single unit is called encapsulation. The programming elements used to manipulate the data in the object are called methods. Encapsulation packs the object's data

and methods together into a single unit. This allows a user to access the data through the available methods with no knowledge of how the data are stored or how the methods perform their tasks. The user simply executes a method that performs some desired action on the data. Objects are further defined by properties, which tell something about the object. The color, the size, and the font used are all examples of properties an object might have. The specific properties associated with an object depend on the object itself and the information needed to define its look, feel, and behavior. Visual Basic 6.0 and earlier versions are not true object-oriented programming languages (Bradley & Millspaugh, 1998; Harriger et al., 1999; Shelly et al., 1999).

A working knowledge of inheritance and polymorphism are crucial to an understanding of objects in the programming environment. Objects are part of object classes, which are plans or specifications used to create multiple objects with the same attributes and behaviors. A specific object of a given class is called an instance of that class. The instance automatically contains all the attributes and behaviors defined in the class. New object classes can be created by modifying existing object classes. Such a new object class is called a subclass. Through a mechanism called inheritance, the new subclass is defined in terms of the difference between itself and the object class from which it originated (Shelly et al., 1999).

Polymorphism is the ability of a method to perform different tasks, depending on the object it is associated with. Polymorphism allows instructions for object manipulation to be given in more general terms. For example, a method called Start can represent a general activity that can be associated with a variety of objects such as automobiles or microwave ovens. Rather than provide specific, detailed instruction on how to start either a car or a microwave oven, the user simply initiates the Start method. The Start method exists in both objects, but it performs a very different activity on the car than on the microwave. The user does not need to know anything about these activities or how they differ. The Start method will perform the appropriate start actions on the object (Lahotka, 2002).

Only Visual Basic.NET is a true object language. Versions prior to .NET use an inheritance mechanism that does not conform to the strict definition of object inheritance.

Visual Programming and Visual Basic

Whether or not Visual Basic is a visual programming language depends on the definition of visual programming applied, visual environment or visual syntax. An application development system that contains graphical tools that aid in creating and manipulating a program written in a text-based language is called a visual environment. Visual Basic is a visual environment programming language. Programs are created by selecting common Windows controls from a graphical list called the Tool Box. The mouse is used to select a control from the list and place it on the form. Once on the form, controls can be moved, resized, and aligned using the mouse. The complete user interface for a program can be created and refined in the Visual Basic environment without typing a single character on the keyboard (Harriger et al., 1999).

An application development system that expresses syntax (or grammar) in a graphical manner via pictures, diagrams, and symbols is called a visual syntax environment. A complete program can be created with no keyboard text entry. The diagrams, symbols, and pictures employed show containment and relationships between the various elements of the program and how these elements should be manipulated. Visual Basic does not fit these criteria. Most program statements in Visual Basic must be entered as text via the keyboard. As statements are entered the development environment will provide some level of context-sensitive help. Pop-up message windows provide syntax and format assistance specific to the language element. Therefore, the more restrictive visual syntax definition of visual programming cannot be applied to Visual Basic (Harriger et al., 1999).

Event-Driven Programming in Visual Basic

Objects have events associated with them. These events are usually actions that can be performed on the objects by an outside source, such as the user. Examples of events in a Windows environment include clicking a mouse button, keyboard input, moving the mouse, changing the focus to a different control, and dragging and dropping. The object's response to a given event is expressed as a series of programming language statements that are contained in a method called an event procedure. This event procedure determines how the object will respond to the event. A large part of the programming task is identifying the events to which the program must respond and coding those responses (Harriger et al., 1999).

Events have ushered in a new approach to programming. Prior to graphical user interfaces (GUIs), all programming was procedural in nature. The program determined what actions would be taken and their order. With the GUI, the events and the order in which they are performed dictate the actions that occur within the program. The ability of the user to perform multiple events on multiple objects in any sequence is completely inconsistent with procedural design (McKeown, 1999). Programs must now be designed without restrictions on the sequence or timing of the user's interaction. The emphasis has shifted from controlling user interaction to reacting to the user interaction. This subtle change has created a completely new event driven approach to application design and development.

THE VISUAL BASIC USER INTERFACE

Getting Started in the Visual Basic Development Environment

Visual Basic offers a completely integrated development environment (IDE) from which a programmer can design the form, enter code, and test, debug, and save a program. The user does not have to leave the integrated development environment. This ability to move quickly between the development and testing stages promotes faster and more efficient application development. Once development is complete, the program can be saved and compiled

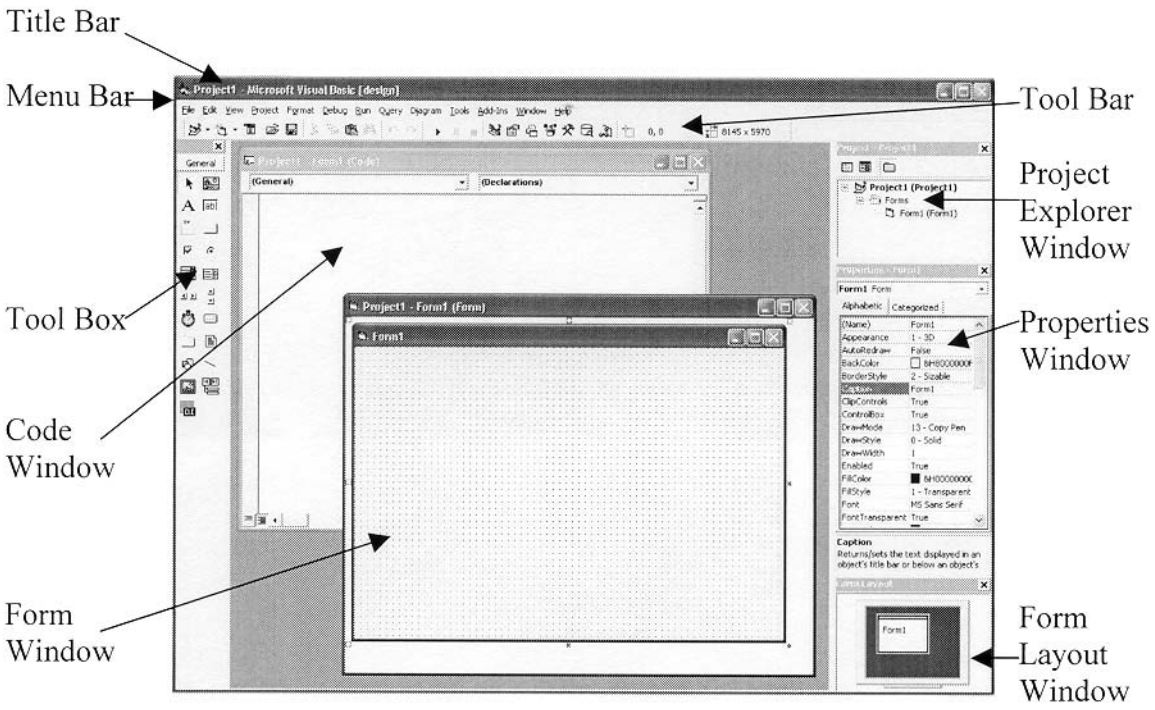


Figure 1: The Visual Basic IDE.

into an executable form, again without leaving the development environment.

The Visual Basic development environment opens by displaying a dialog box containing three tabs: New, Existing, and Recent. The New tab offers a wide variety of templates that can be used to create different types of applications. These application types include the following:

Standard.EXE files: normal Visual Basic programs.

ActiveX.EXE files: programs that allow other programs to access their data.

ActiveX.DLL files (Dynamic Link Libraries): collections of functions and procedures which can be accessed by other programs.

ActiveX Controls: controls that can be incorporated into other programs.

Active Document.DLL and.EXE files: programs that can work within a browser and can be distributed over the Internet or an intranet.

Add-In files: files that can be added to the Visual Basic programming environment to automate common and repetitive tasks.

The Visual Basic Application Interface assists in constructing the user interface.

The Existing tab presents a standard Windows dialog box for browsing through local and network folders. An existing Visual Basic program can be selected for modification. The Recent tab lists programs that have been recently created or modified (Harrington et al., 1997).

After a new or existing program has been selected, the Visual Basic IDE appears. The IDE consists of several panes that provide general information about the program being developed. A workspace to enter and edit code is also included. See Figure 1.

Form Layout Window

The Form Layout window is used to design the user interface. Think of a form as a blank window ready to be filled with controls. Visual Basic programs can contain more than one window. Each window is created in a separate form (Harrington et al., 1997). Controls from the Toolbox are placed on the form by clicking the control, moving onto the form, and dragging a box open. Releasing the mouse button places the selected control on the form inside the drag box. The control is automatically sized to fill the box. See *Creating a Program in the Visual Basic Integrated Development Environment* section later in this document for specific examples of placing controls on a form.

Code Window

Placing the controls on the form merely creates the interface. Blocks of Visual Basic statements, called code, must be added to the appropriate event handling procedures to make the program respond to the event properly. Visual Basic code is entered in the Code window.

Toolbox

The Toolbox displays icons that represent the various controls that can be placed on a form. The Toolbox does not automatically display all possible controls. Additional controls are available from within the Visual Basic environment and can also be obtained from third-party vendors. These additional controls provide greater functionality, which can decrease development time (McKeown, 1999).

Menu Bar
The Menu bar control is common to most Windows programs. A menu of related commands drops down when an item on the Menu bar is clicked. These commands are also invoked by clicking. Some menu items initiate an action immediately, whereas others open additional menus or dialog boxes.

Toolbar
The Toolbar control is also common in Windows programs. It provides icon-based access to the items in the menus. In addition to the Standard Toolbar, Visual Basic provides several specialized Tool bars for specific tasks such as drawing. Click View on the Menu bar to see a complete list of the available toolbars.

Project Explorer
The complete Visual Basic program is called a Project. The Project Explorer window uses a hierarchical format to list all forms, procedure modules, and objects in the program. The Project Explorer allows the programmer to instantly access any part of the project by clicking on an item in the hierarchical list.

Properties Window
All objects have sets of properties that define how the objects will appear and behave. Some properties can be assigned values when a program is created, whereas others are assigned values by the computer system while the program executes. All properties that can be set by the programmer are listed in the Properties window. Each property is listed in the left column, with its current value displayed in the column to its right. These values can be changed by clicking the value in the right column and entering a new value in its place. Property values can also be changed during program execution by including appropriate programming statements. See *Creating a Program in the Visual Basic Integrated Development Environment* later in this document for specific examples of setting property values. Initially setting the Caption property to nothing illustrates how to change a property value in the Properties window. The Visual Basic statement *Label1.Caption = "Hello!"* is an example of changing a property value with programming statements.

Form Layout Window
The Form Layout window shows the relative position of the form on the screen during program execution.

Immediate Window
The Immediate window, sometimes called the Debug window, is not present when Visual Basic is first started. The Immediate window appears at the bottom of the screen below the Code and the Form Layout windows. When the program executes from within the development environment the Immediate window displays explanations of errors encountered. The information displayed in the Immediate window, when used in conjunction with the other debugging tools available in Visual Basic, reduces development time by speeding error identification and correction.

Creating a Program in the Visual Basic Integrated Development Environment
The following example will help illustrate the various components of Visual Basic's IDE. The task for this example is to display a simple greeting when the user clicks a button. Start Visual Basic and select Standard EXE from the New Project dialog box. Begin creating the user interface by placing a Command Button on the form. This Command Button, when clicked, will cause the greeting to appear on the screen. Click the Command Button icon on the Toolbox, move to the lower center of the Form Layout Window, and drag open a rectangular box. The Command Button will appear when the mouse button is released. With the Command Button still selected, click Caption in the Properties Window. Change the value in the right column to "Greet." This will cause the text on the button to be changed to "Greet."

Next place a label on the form to display the greeting. Select the Label icon on the Toolbox and drag open a rectangular box in the upper center of the Form Layout Window. The Label will appear when the mouse button is released. Change the value in the Caption property of the label to display nothing by erasing all characters in the right column. The form should resemble Figure 2.

The interface is now complete. The controls, which are objects, have inherited the operating characteristics and appearance of similar Windows controls. Test the interface by running the program. Click the Run symbol, the right-pointing arrow in the Toolbar just below the Diagram menu in the Menu Bar. The Visual Basic Development Environment will run the program. Try clicking the "Greet" button. It will act like a normal Windows button, but the greeting will not display when the button is clicked. Clicking is an event that a user can perform on the "Greet" command button. To display the greeting, code must be entered into the event-handling procedure to control what happens when the button is clicked.

Stop the program by clicking the Stop symbol in the Tool Bar. It is the icon just below the Add-Ins menu in the Menu Bar. Double click the "Greet" button on the form. The view will shift from the Form Layout Window to the Code Window. Visual Basic automatically enters the header and footer needed to designate this as the procedure that defines what happens when the "Greet" button is clicked. Between the procedure header, *Private Sub Command1_Click()*, and the procedure footer, *End Sub*, enter the text *Label1.Caption="Hello".*

```
Private Sub Command1_Click()

      Label1.Caption="Hello!"

End Sub
```

This command will change the Caption property of the Label control named Label1 to "Hello!" Run the program again (click the Run symbol) and click the Greet Button. The label will change from blank to "Hello!." Stop the program (click the Stop symbol). Save the program by selecting the Save option from the File menu on the Menu Bar. Following the instructions in the subsequent dialog boxes to complete the save operation.

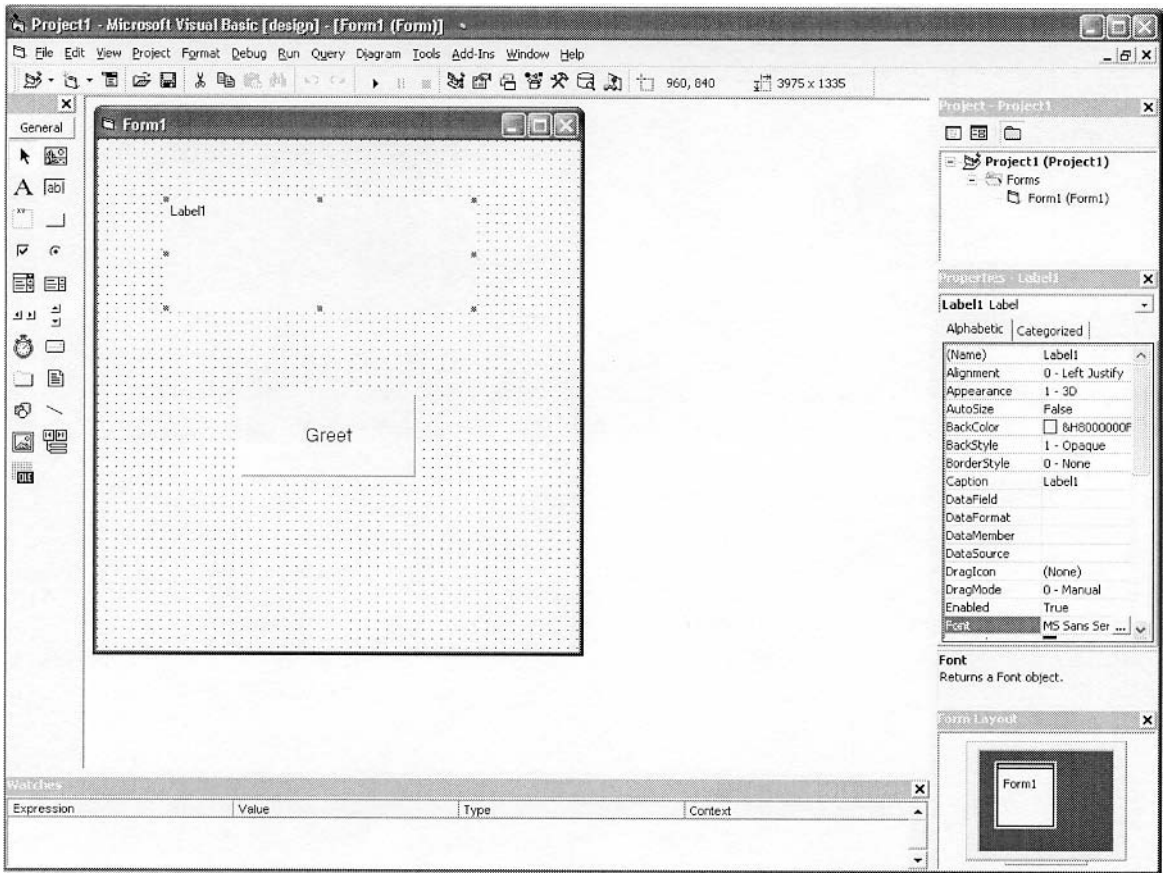


Figure 2: Greet Exercise sample form.

Program Development and Debugging

Programming errors happen. Any program of any significant size or complexity will certainly have at least one error in it. The process of locating and fixing these errors is called debugging. Since debugging is such an important (and frequent) part of the application development process, Visual Basic includes special tools to aid in locating and correcting problems

Error Types

There are three types of programming errors: syntax errors, runtime errors, and logic errors. Syntax is the set of rules that must be followed when language statements are written. Syntax errors occur when program statements do not follow these rules. Misspelling a Visual Basic keyword is an example of a syntax error.

Runtime errors occur when the computer cannot execute a statement. The statement is syntactically correct, but some aspect of the command prevents it from being executed. For example, the statement $C = A/B$ is syntactically correct. However, if B has a value of zero at runtime, the statement cannot be executed because division by zero is not mathematically possible.

The final type of error is the logic error. All statements in the program are syntactically correct and the program encounters no runtime errors, but the results are not as expected. The instructions are perfectly legitimate Visual Basic, but they do not properly perform the specified task. An empty label on a form where output was expected is an example of a logic error (Eliason & Malarkey, 1998).

Visual Basic Debugging Tools

One of the most useful Visual Basic debugging tools is the breakpoint. A breakpoint is a marker placed in the program. The program will temporarily stop running when a breakpoint is encountered. This allows the programmer to use other debugging tools to obtain information about the program while it is paused in midexecution. To set a breakpoint, first place the cursor on the desired program statement. Press F9 or click in the gray margin to the left of the statement. The statement will be highlighted, usually in red, and a large dot will appear in the grey margin to the left of the line. At execution, the program pauses before executing this statement. The Visual Basic environment is now in Break mode. The statement at which the program paused is highlighted and a yellow arrow appears on the dot in the gray margin. To continue executing the program, click on the Run symbol in the Standard toolbar. The program will continue to execute until it reaches another breakpoint or runs to completion. Breakpoints are removed by clicking the dot in the gray margin to the left of the statement.

Once the program reaches a breakpoint and pauses, the programmer can check the values in selected variables by using the Watch window. Selecting a variable to be "watched" opens the Watch window. The current

values of all variables being watched are displayed in the Watch window. Variables can be added to the Watch window before the program is executed, or when a breakpoint is reached. To add a variable to the Watch list, right click the variable name in a program statement and select the Add Watch option. To remove a variable from the Watch list, select the variable in the Watch window and click the Debug menu in the Menu bar. Select Edit Watch and click Delete (Eliason & Malarkey, 1998).

Another very useful Visual Basic debugging tool is the Data Tips window. While the program is in break mode, position the cursor over a variable name in the code. A Data Tips window will display showing the current contents of the variable. Data Tips windows are used to check the contents of variables not in the Watch window (Harriger et al., 1999).

Visual Basic can also step through the program, executing one instruction at a time. This is called Single Stepping. After executing an instruction, the program highlights the next instruction and pauses in the same manner as at a breakpoint. Pressing the F8 key causes the highlight instruction to execute. The next instruction is highlighted and execution pauses again. A breakpoint is usually set to stop program execution at a specified point. The Single Step feature allows execution of all or part of the remaining instructions one at a time (Harriger et al., 1999).

The debugging tools available in Visual Basic are most effective when used in combination. Watch windows and Data Tips windows allow the programmer to see what is happening "inside" the program as it runs. This information can help identify instructions that are not producing expected results. Single Stepping through a program also allows the programmer to see exactly which statements are executing and in what order. This can help identify problems in modules, loops, and decision structures.

VISUAL BASIC LANGUAGE ELEMENTS

Visual Basic is a high-level programming language with a rich command set. Its heritage as an instructional language makes Visual Basic suitable for a wide range of programming tasks. Over the years, vendor refinements and enhancements have addressed and corrected most of the problems and shortfalls of the original Dartmouth specifications. Some key elements are discussed below.

Variables

Specific variable declarations are not required in the original BASIC language. This option remained in Visual Basic through Version 6.0 (Evjen & Beres, 2002). Although using variables that are not specifically declared may speed program development somewhat, they are difficult to manage and often lead to errors. Any variable not specifically declared is created with the Variant data type, which is very inefficient. Good programming practice dictates that variables be specifically declared before use. To facilitate this, Visual Basic through Version 6.0 offered the Option Explicit statement. The Option Explicit statement is placed in the General Declarations section of the Visual Basic program. It disables Visual Basic's automatic variable declaration option, forcing all variables to be declared in the program's code. Automatic variable declaration is not supported in the latest release of Visual Basic, Visual Basic .NET (Evjen & Beres, 2002).

Variables are declared using the Dim statement. As with all programming languages, each variable must have a unique name. With the release of Visual Basic.NET, all data types follow the Common Language Specification (CLS) incorporated into Microsoft's .NET framework. To achieve this compliance, some of the data types available in Visual Basic 6.0 and earlier versions have been eliminated. Other Visual Basic 6.0 and earlier data types have been changed to fit CLS naming conventions. Below is a partial list of data types available in Visual Basic. Differences between Visual Basic 6.0 and Visual Basic.NET are noted (Evjen & Beres, 2002).

Boolean

The Boolean data type is a 2-byte field that contains True or False values.

Char

Char is a 2-byte field that contains unsigned integer values between 0 and 65,535.

Date

Date is an 8-byte field that contains date values between January 1, 1 and December 31, 9999. The Date type in Visual Basic 6.0 and earlier versions is replaced by the Long data type in Visual Basic.NET.

Decimal

Decimal is a 12-byte field that contains real and integer numbers. The integer range is ±79,228,162,514,264,337,539,950,335. The real number range is ±7.9228162512265337593543950335 with 28 places to the right of the decimal point. This data type is called Currency in Visual Basic 6.0 and earlier versions.

Double

Double is an 8-byte field that contains real numbers in the range from −1.797693134862231E308 to 1.797693134862231E308.

Integer

Integer is a 4-byte field that contains integer values in the range ±2,147,483,648. The Integer data type in Visual Basic 6.0 and earlier versions is equivalent to Visual Basic.NET's Short data type.

Long

Long is an 8-byte field that contains integer values in the range −9,223,372,036,854,775,808 to 9,223,372,036,854,775,807. This data type did not exist in Visual Basic versions prior to Visual Basic.NET.

Short

Short is a 2-byte field that contains integer values between −32,768 and 32,767. This data type was called Integer in Visual Basic 6.0 and earlier versions.

String
String is a field of at least 12 bytes and a maximum limited only by the computer hardware. It is capable of holding from 0 to approximately 2 billion Unicode characters.

Variant
The variant data type available in Visual Basic 6.0 and earlier versions does not exist in Visual Basic.NET (Evjen & Beres, 2002). Variables declared in the Variant data type can hold integers, real numbers, or character strings. Variables created via the automatic variable declaration mechanism are of the data type Variant. Elimination of the Variant data type from Visual Basic.NET made support of the automatic variable declaration option impossible. Given the Variant data type's inefficient use of memory, the additional overhead needed to actively configure memory usage to accommodate both numeric and character string data, and the development problems discussed above, elimination of the Variant data type and the automatic variable declaration option from Visual Basic.NET is not considered a significant loss.

Control Structures
Visual Basic offers a complete set of decision and repetition control structures. Decision structures allow the program to compare data items to determine which statements to execute. Visual Basic includes the If, If . . . Else, and Select Case decision control structures. Repetition structures allow the program to repeat groups of statements, called blocks, if certain conditions are met. Visual Basic includes the pretest While loop, the pretest Until loop, the posttest While loop, the posttest Until loop, and the Counter Controlled loop. These control structures are not unique to Visual Basic. Consult the bibliography for text and reference books that cover the general function and specific syntax of Visual Basic's control structures.

Common Controls
The Toolbox contains the control classes that can be included in a Visual Basic program. It contains controls commonly used in Windows programs. Additional controls are available from other Microsoft sources and from third party vendors. A control is selected from the toolbox and placed on the form. Each control is given a unique name. Controls in Visual Basic are object classes, which have properties and methods defined for them. These properties and methods are inherited by the control when it is created. Properties and methods associated with a specific control are accessed using the . (read "dot") operator. The . operator is used to separate the control name from the property or method being accessed. For example,

```
Label1.Caption
```

will access the Caption property of the Label control called Label1. A value can be assigned to this property in the same manner in which a value is assigned to a variable.

```
Label1.Caption="Hello!"
```

The following is a brief description of some of the more commonly used controls. Unless specified otherwise, these controls are available through the standard Toolbox.

Form
The Form can be considered a control. It has properties and methods associated with it, but it is not on the Toolbox. A form can be added to a project by selecting the Add Item button on the Tool Bar, or from the Project menu in the Menu Bar (Harriger et al., 1999).

Label Control
The Label control is used to display text. It is often used to identify the function of some other control, to display information generated by the program, and to display error messages (Harriger et al., 1999). Through manipulation of the Label's properties, such attributes as the size and shape of the display area and the font, the color, and the number of lines of text displayed in the label can be specified.

Text Box Control
The Text Box control is used to enter data. Text Box controls inherit full editing capabilities on entered text. The user can type, backspace, delete, and select text using common Windows editing techniques (Harrington et al., 1997).

Command Button Control
The Command Button allows the user to initiate some action. This action is usually initiated by clicking the command button (Harriger et al., 1999).

Check Box Control
The Check Box control allows the user to select one or more items from a list. By clicking the small box next to the description of the item, the user can insert or remove a check mark. Instructions placed in the click-event-handling module for the Check Box control determine which boxes are checked and therefore which items the user selected (Harrington et al., 1997).

Option Button Control
The Option Button control functions similar to the Check Box control, except that only one of the items in the list can be selected. Selecting one item automatically deselects all other items. Option Buttons are sometimes called radio buttons because they function like the station selection buttons on a car radio. Only one station can be selected at any one time. Option Buttons are frequently used with the Frame Control. The Frame control groups the Option Buttons together and specifies which Option Buttons work together. This allows two or more groups of Option Buttons to operate independently on the same form (Harrington et al., 1977).

List Box Control
The List Box control is used to display a list of items. Items may be added to and removed from the list via the AddItem and RemoveItem methods, respectively. The user may select an item from the list. The Text property of the

List Box contains the selected item. The List Box can be emptied by using the Clear method.

Combo Box Control

The Combo Box control is a combination of the List Box and Text Box controls. The Combo Box is a List Box with a Text Box affixed to the top. The Text Box portion remains visible whereas the List Box portion can be displayed or hidden. An item selected from the Combo Box list is displayed in the Text Box. In many applications, the user does not need to see the contents of the list once an item has been selected. The Combo Box displays the complete list while a selection is made, then hides the list to save space on the form. The Text Box portion of the Combo Box remains visible, keeping the selected item visible on the form.

Modularity: Functions and Procedures

The sheer size and scope of business applications require that they be broken down into smaller, more manageable units. Each unit contains a relatively small block of code that addresses one task or aspect of the overall program. Each event in an event-driven programming environment must have a separate unit of code associated with it that defines the actions taken when the event occurs. This critical concept, called modularity, is fundamental to all high-level programming languages. Modularization helps eliminate duplicate code, improves program understandability, and facilitates reusability (Harriger et al., 1999). Visual Basic supports modularity through procedures and functions. One or more data items can be transferred to the module when it is executed. These data items are called *parameters* or *arguments*.

The act of transferring program execution to a module is known as calling the module. The statement that sends program execution to the module is the Call statement. The module that contains the Call statement is the Calling routine. The module that is executed is the Called routine. Sending parameters to a module is called passing parameters. Parameters are passed to a module by listing them in parentheses immediately after the call.

Functions and procedures differ in how the program behaves after the module executes. Procedures return control to the calling point and sequential execution continues. Functions return control to the calling point, along with a single data item. This returned data item replaces the function in the calling statement and sequential execution continues.

Procedures use the Visual Basic keyword Call to execute a module. The Call keyword is followed by the module's name. The parameter list enclosed in parentheses immediately follows the module name. Functions do not use a specific Visual Basic keyword to initiate the call. The function name, followed by the parameter list enclosed in parentheses, is placed in a statement where a constant or variable could be located. When the function name is encountered, the module is executed. The value returned by the function is substituted back into the calling statement and the program continues to evaluate that statement.

Visual Basic Function Library

Visual Basic provides an extensive library of prewritten functions. These functions handle common tasks such as mathematical calculation and character string manipulation. Unlike some languages, no special linking is needed to make function libraries available to a Visual Basic program. Library functions are called in the same manner as user-defined functions.

DATA ACCESS

Files

Visual Basic supports sequential and random access files, although random access files are not widely used today. Data manipulation needs that had been traditionally met by random access files are now performed by database systems. Currently, files are used to store relatively small amounts of data that require simple processing. One of the four sequential file formats supported by Visual Basic can usually satisfy these modest requirements.

Visual Basic refers to files by numbers that are assigned when the files are opened. File handling statements incorporate this file number identification system to differentiate between open files. All file types supported by Visual Basic use common open and close statements.

Comma-Separated Values

Comma-Separated Values (CSV) is a commonly used file format in which individual fields within the record are separated by commas. String data are enclosed in double quotation marks. CSV files contain variable-length records; the number of characters in each record does not have to be the same.

Report–Record

Report–Record files are designed to be viewed, usually in a printed form. Extra formatting data are included in each record to enhance its appearance. Data fields are the same length and align in vertical columns when viewed in a report format. Pagination, page headers, page footers, and column headers may also be included.

Fixed Record-Length Type 1 and Type 2 Formats

Fixed Record-Length Type 1 files are compatible with files used by programs written in the COBOL programming language. Although little new development is being done in COBOL, a large body of legacy COBOL still exists. All field lengths are fixed, so all records are the same length. Character and numeric data are stored in ANSI character format (Harriger et al., 1999). Fixed Record-Length Type 2 files are similar to Type 1 files, with the exception that numeric data are stored in binary format.

Database and Data Access

Visual Basic's data access features provide a convenient way to create powerful front-end programs for database access. These front-end programs insulate end users from complex data management systems. Front-end programs also provide a measure of data security, allowing control over the data accessed and the types of operations allowed. Visual Basic 6.0 supports Microsoft's Access

database, as well as dBASE III, dBASE IV, dBASE 5.0, Excel, FoxPro, Lotus, Paradox, and text file formats. With the addition of Open Data Base Connectivity, Visual Basic can access a wider range of database systems such as SQL Server, Oracle, and DB2 (Bradley & Millspaugh, 1998).

With the release of Visual Basic.NET, all data access is controlled through the .NET framework. The data are combined with the programs needed to access and manipulate the data. This data and programs unit is then treated as an object. The data corresponds to the data component of an object and the programs correspond the methods component of an object.

VISUAL BASIC.NET—VISUAL BASIC FOR INTERNET APPLICATIONS

Visual Basic through Version 6.0 has established itself as a significant player in Windows programming. As computer information systems evolve toward Web and Internet technology-based data delivery, Visual Basic has evolved with them. Microsoft has taken a significant step in this evolution with the release of Visual Basic.NET. Understanding the differences between Visual Basic.NET and previous Visual Basic releases requires an understanding of Microsoft's .NET framework.

Simply stated, .NET is Microsoft's platform for building XML Web Services. The goal of .NET is to provide a platform, or framework, for developing and deploying Web-based data delivery services, called Web Services, in a simple, secure, and consistent manner. Although Web Services are the central thrust of .NET, Visual Basic.NET can be used for much more than building applications that create and consume Web Services (Evjen & Beres, 2002).

Software has become a service in today's computing environment, prompting a paradigm shift. Buying a shrink-wrapped software package off the store shelf is giving way to buying software services that are delivered over the Internet or Internet-like connections (Evjen & Beres, 2002). Instead of a box of CDs or DVDs, software will be purchased via the Internet and downloaded directly into the computer when needed. The software service will provide support and product updates via the same Internet mechanism. Creating these Web Services is the next step in application development evolution. For the Internet to be an effective and efficient data delivery mechanism, a common communications method, or protocol, is needed. This protocol must support a wide variety of transactions without regard to the hardware, operating system, or browser employed. Hypertext transfer protocol, or HTTP, has existed for many years and is an established Internet data exchange protocol. It is well suited to data delivery across the Internet. The simple object access protocol, SOAP, was the first attempt at a standard protocol to move data over the HTTP transaction transport mechanism. SOAP is an XML-based specification for sending and receiving data over the Internet. XML is a self-describing text file that can be understood by any operating system or browser. Using the SOAP specification to format data and the HTTP protocol to transport them, any Web site can offer data-related services. SOAP formatted requests are sent to the server via the HTTP protocol. The server responds via XML within the SOAP framework, using HTTP to deliver the XML response. XML's broad application across many different hardware and software platforms will give the .NET initiative great appeal to a wide range of end users. Simple handheld devices capable of running browser software will rival laptop and desktop computer systems in data access capabilities. This should further speed .NET's acceptance (Evjen & Beres, 2002).

To the developer, the .NET initiative has different ramifications. All application development will now occur on the server side of the Internet. Applications will no longer be installed on client computers. The client will run only a browser. Applications will either run on the server and push data to a browser, or download from the server and execute in a browser. Applications must be specifically designed with this Internet delivery in mind. If design and development of .NET-ready applications is difficult and costly, .NET will never gain acceptance. Anticipating this, Microsoft created several tools to aid in .NET application development. Visual Studio.NET, which contains Visual Basic.NET, provides programming tools that can easily and quickly create Web Services.

XML is such a basic component of .NET that a mastery of it would seem to be essential to successful .NET development. Actually, the XML portions of .NET are buried in the .NET framework. Application developers use tools like Visual Studio.NET, which automatically generate needed XML. .NET programmers can create Web Services with no knowledge of XML. This makes Visual Basic.NET and the other languages in Visual Studio.NET true RAD tools on both the client and server sides (Evjen & Beres, 2002).

The .NET Framework

The .NET framework is a collection of components needed to create Web Services. Figure 3 shows the key .NET framework components (Evjen & Beres, 2002).

Of specific interest are the common language Runtime and its base class library. A basic understanding of assemblies and name spaces is also useful in understanding .NET programming (Aitken & Syme, 2002).

The Common Language Runtime

There are currently approximately 18 languages that support the .NET initiative (Evjen & Beres, 2002). All .NET applications, regardless of the language they were

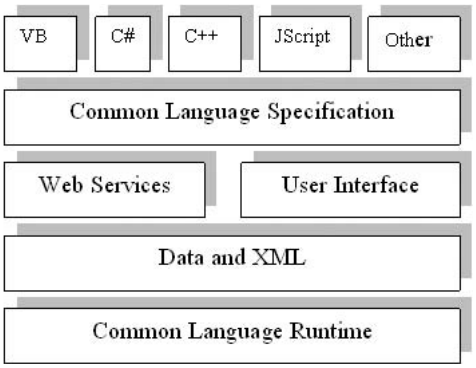


Figure 3: .NET framework.

originally written in, execute through the common language Runtime (CLR). The base class library is a library of object classes that are available to all .NET languages. Current literature disagrees on whether the base class library is a separate element of the framework or part of the CLR (Aitken & Syme, 2002; Evjen & Beres, 2002). Some of the base class library deals with non-Internet-related issues. Windows NT event log manipulation is one example. Others classes are specific to Internet programming and Web Services development. An application written in a .NET language is compiled into an "intermediate language" (IL). All .NET applications compile to this common IL, regardless of what language is used to create the program. Developers never deal directly with the IL version of the program; IL is the output of an intermediate compilation step. Therefore, the actual makeup of the IL is not important to a basic understanding of .NET. It is sufficient to know that IL is optimized for quick compilation into machine-specific code. When a Web server calls the .NET application, the IL code is passed to the common language Runtime, which compiles the IL code to machine-specific code and executes the program. Each module is compiled on an as-needed basis. Modules that are not called during execution are not compiled. Compiled functions are cached for future reference (Aitken & Syme, 2002; Evjen & Beres, 2002).

This dual compilation mechanism (source code to IL and IL to CLR) seems complex and would appear to slow execution, but it actually has several advantages. Since only executed modules are compiled, less memory is needed at runtime. The CLR is the only machine-specific portion of the .NET framework. Porting the framework to another computing platform requires only that the CLR be changed. Applications require no changes. At runtime, the CLR translates the IL code into platform-specific machine language. Any hardware platform that supports the .NET framework can run applications developed on any platform within the framework. The base class library standardizes the set of object classes between all applications and hardware platforms.

Assemblies

.NET applications are deployed in assemblies. An assembly is a combination of IL code, machine language code, and a manifest (Evjen & Beres, 2002). The manifest contains the name of the assembly, version, information on all object classes used, and other data the CLR needs to run the application (Aitken & Syme, 2002). The manifest is metadata, information about an entity that is contained in the entity itself. The use of metadata eliminates the need to use the system registry to associate DLL files with an application. To deploy a .NET application, simply copy the assembly into the appropriate folder. Since all the information needed to execute the assembly is contained in the assembly, any calling routine need only be able to find the assembly to execute it. Complex installations requiring changes to the system registry are a thing of the past with .NET.

Namespaces

A namespace is a repository, or container, for a set of object classes. Using namespaces allows two object classes of the same name to coexist simultaneously, each in its own namespace (Aitken & Syme, 2002). This concept roughly equates to the scope property of variables. Two variables of the same name may exist in the same program provided they are created in different procedures. Two classes of the same name may exist in a single system if they are placed in different namespaces. Namespaces allow programmers to develop classes without regard to other classes in the system. This facilitates the team approach to application development and allows assemblies to be incorporated into other programs without conflicts.

Visual Basic.NET

Visual Basic.NET and the .NET framework offer significantly enhanced tools to speed application development. Visual Basic.NET language innovations, RAD features, new forms models for Web-based applications, and new forms models for Windows-based applications create a powerful application development environment. These enhancements have resulted in some significant changes between Visual Basic.NET and previous versions of Visual Basic.

Language Innovations

Visual Basic.NET is now a fully object-oriented language. Previous versions of Visual Basic did not fully implement all aspects of objects, specifically with respect to inheritance. Visual Basic.NET fully supports inheritance, polymorphism, encapsulation, overloading, and overriding (Lhotka, 2002). Visual Basic.NET incorporates an object-based exception handling system that provides a consistent means of dealing with errors. Errors involving methods, multiple method chains, and external components including those created in other .NET languages are all handled through the same error object, eliminating separate error handlers for each source.

RAD Features

Visual Basic.NET includes most of the features that make Version 6.0 a RAD environment. With .NET, these RAD tools can be applied to client- and server-based application development. The Visual Basic.NET integrated design environment (IDE) is similar to Visual Basic 6.0. Developers already familiar with other versions of Visual Basic or Visual InterDev can quickly transition to Visual Basic.NET.

Web Forms

The inclusion of Web forms allows development of Web-based applications in the familiar Visual Basic environment. Developers do not need to learn a different IDE. All tools and features available for traditional Windows-based application design are available for Web-based applications design.

Web Services

Creating Web Services in Visual Basic.NET is as simple as adding the <webmethod> identifier to a method. The underlying code that enables the Web Service is created and managed by Visual Basic.NET's IDE and .NET's underlying framework.

Windows Forms

Conventional Windows forms have been revised to incorporate the new .NET framework. Visual Basic.NET's use of the base class library allows all Windows forms to run inside the browser, eliminating the need for applets (Evjen & Beres, 2002).

Visual Basic 6.0 and Visual Basic.NET Differences

Many of the changes in Visual Basic.NET revolve around the underlying .NET framework. Some elements of Visual Basic 6.0 are not compatible with the .NET framework. A complete listing of the differences between Visual Basic 6.0 and Visual Basic.NET is well beyond the scope of this work. Many good references are available on this topic. The author recommends the *Visual Basic.NET Bible* listed in the References.

The use of the common language Runtime and the base class library have forced all .NET languages to use common data types, control structures, function calls, variable scoping, and variable declarations. Many Visual Basic 6.0 elements exist in the Visual Basic.NET. Others, like data type definitions, required modification to be consistent with .NET. Some aspects of the original BASIC language cannot be supported in .NET. For example, BASIC's While . . . Wend structure and the variant data type are inconsistent with the base class library and are no longer supported.

CONCLUSION

Visual Basic has been a popular RAD tool for many years. It provides a programmer-friendly IDE that speeds form layout and application debugging. Visual Basic's evolution from the popular BASIC language has further contributed to its wide use. Microsoft's industry prominence has certainly furthered Visual Basic's acceptance. With the introduction of .NET Microsoft is positioned to take a leading role in the Web Services market, and Visual Basic.NET is its main application development tool. The success of Visual Basic.NET will depend on the acceptance of the .NET initiative and Visual Basic.NET's ability to continue as a solid Windows application development tool. Some experts feel that .NET brings nothing new to Web programming and, therefore, will not have a major impact (Petreley, 2002). Others do not see Visual Basic.NET as a significant improvement and question the practicality of converting current Visual Basic 6.0 applications (Berger, 2002; Evjen & Beres, 2002). Clearly, the future of the .NET initiative will drive Visual Basic.NET's role in Web Services development. Regardless of Visual Basic.NET's success as a Web Services development tool, maintenance of legacy Visual Basic systems alone will keep Visual Basic as an industry standard for many years to come.

GLOSSARY

Common Language Specification The .NET framework requires that all programs compile to a common intermediate language. To accomplish this, all languages must share common data types, control structures, and modularization schemes.

High-level language A programming language that compiles a single statement into many machine language statements. High-level languages support structured programming.

Integrated Development Environment A program that contains a complete set of tools to create, debug, test, and save a program. All these features are available within a single environment.

Intranet A private network based on Internet and World Wide Web technologies.

Variable length records Certain file formats do not require that corresponding fields in each record contain the same number of characters. Therefore, some records will contain more characters than others.

Web applications Programs designed to run across the World Wide Web.

Windows applications Programs designed to run in a Microsoft Windows operating system.

CROSS REFERENCES

See *Visual Basic Scripting Edition (VBScript); Visual C++ (Microsoft)*.

REFERENCES

Aitken, P., & Syme, P. (2002). *Visual Basic.NET web programming in 21 days*. Indiana: Sams Publishing.

Berger, M. (2002). Microsoft cautions against quick move to .NET. *Infoworld*. Retrieved April 24, 2002, from http://staging.infoworld.com/articles/hn/xml/02/02/15/020215hmmigrate.xml

Bradley, J. C., & Millspaugh, A. C. (1998). *Programming in Visual Basic*. Massachusetts: Irwin/McGraw-Hill.

Eliason, A., & Malarkey, R. (1998). *Visual Basic 5*. Indiana: Que Education & Training.

Evjen, B., & Beres, J. (2002). *Visual Basic.NET bible*. New York: Hungry Minds, Inc.

Hare, V. C., Jr. (1982). *BASIC programming* (2nd ed.), New York: Harcourt Brace Jovanovich.

Harriger, A. R., Lisack, S. K., Gotwals, J. K., & Lutes, K. D. (1999). *Introduction to programming with Visual Basic 6: A problem solving approach*. Indiana: Que.

Harrington, J., Spenik, M., Brumbaugh, H., & Diamond, C. (1997). *Visual Basic 5 interactive course*. California: The Waite Group, Inc.

Lahotka, R. (2002). Visual Basic.NET as a fully object-oriented language. Retrieved April 15, 2002, from http://softwaredev.earthweb.com/vb/print/0,12080_882111,00.html

McKeown, P. G. (1999). *Learning to program with Visual Basic*. New York: J. Wiley.

Petreley, N. (2002). No .NET advantage. *Computerworld*. Retrieved from http://www.computerworld.com/itresources/rcstory/0,4167,STO68723_KEY11,00.html

Shelly, G. B., Cashman, T. J., Repede, J. F., & Mick, M. L. (1999). *Microsoft Visual Basic 6: Complete Concepts and techniques*. Massachusetts: Course Technology.

Visual Basic Scripting Edition (VBScript)

Timothy W. Cole, *University of Illinois at Urbana-Champaign*

INTRODUCTION

What VBScript Is and Where It Comes From

BASIC, an acronym for beginner's all-purpose instruction code, was developed in 1964 at Dartmouth College by John Kemeny and Thomas Kurtz. Less powerful, but simpler and easier to learn than other high-level programming languages extant at the time, BASIC was initially used as a teaching tool for introducing undergraduates to the fundamentals of programming and programming languages. Original BASIC lacks many flow control features of newer, structured programming languages and relies heavily on the use of "Go To" statements, a practice that tends to adversely impact code brevity and clarity and that has since been eschewed in more modern languages such as C++ and Java. Original BASIC does not provide good support for modular authoring of applications (e.g., does not provide a way to distinguish between local and global variables and does not invoke external object classes). When ported to the PC, BASIC was initially implemented as an interpreted language (as opposed to a compiled language), meaning that code was processed and interpreted sequentially line-by-line at run time rather than precompiled into an optimized machine-language executable or other intermediate low-level language form at some point in time prior to first execution. This tended to adversely impact performance.

BASIC language interpreters were the first commercial products produced by the Microsoft Corporation, and Microsoft BASIC language implementations were integral components of IBM DOS and Microsoft PC DOS when introduced circa 1980. By the mid-1980s Microsoft and others had significantly enhanced BASIC, introducing many structured programming concepts into the language and making it more powerful and useful for real-world application building. In 1985 Microsoft introduced QuickBasic, which included a BASIC compiler, making for better performance of applications written in the language.

By 1988, as it was developing OS/2 and Microsoft Windows version 3.1, Microsoft had begun work on a Windows-style graphical user interface (GUI) for BASIC code authoring and development. It was at this time that Microsoft purchased from Alan Cooper the rights to Tripod, an early prototype drag-and-drop shell well suited for code authoring in the Windows GUI environment. Visual Basic version 1.0 was publicly introduced in 1991 at Windows World. Visual Basic also marked the initial introduction into the language of several object-oriented programming (OOP) concepts borrowed from C++ and other OOP languages of the day. The object-oriented nature of Visual Basic has continued to evolve since its introduction to the point where Visual Basic.NET (effectively Visual Basic version 7) can now be thought of as a fully OOP language. Microsoft Access BASIC, a precursor of Visual Basic for Applications (VBA), first appeared in 1992, only a year after the introduction of Visual Basic. By the release of Office 97, VBA had become the common language used across Microsoft's primary suite of office productivity applications (e.g., Microsoft Word, Excel, and Access) and by a number of third party applications in lieu of or in addition to application-specific macro languages.

Microsoft Visual Basic Scripting Edition (VBScript), initially a subset of VBA, appeared with the 3.0 version release of Microsoft Internet Explorer in 1996. VBScript was intended in this context as an alternative to JavaScript, support for which previously had been introduced as part of the Netscape Navigator Web browser. VBScript support was later added to Microsoft Internet Information Server (IIS), to Microsoft Outlook, and to the Windows operating system itself (in the form of the Windows Scripting Host [WSH]). This latter implementation allows programmers to create VBScript code to perform functions similar to

those performed using batch files in DOS and shell command scripts in UNIX.

VBScript is today most similar to VBA, but it is not the same. VBScript is a return to an interpreted form of the language, and like VBA, code written in VBScript is designed to be processed and run in the context of another application. The applications supporting VBScript provide application-specific object models to give script authors access to application functions, features, and services. Unlike most VBA implementations, VBScript implementations do not come with a built-in GUI authoring environment. VBScript applications are most often created using simple word processing applications (e.g., Microsoft Notepad) or using multipurpose applications such as Microsoft Visual InterDev. VBA also retains several Visual Basic features not available in VBScript, most notably optional strong data typing, more extensive error handling features, early binding options, and direct access to certain low-level Windows objects and features such as the clipboard and dynamic data exchange functions. The latest releases of VBScript do include support for regular expressions, something not yet intrinsic to VBA.

The Role of VBScript on the Internet and Its Relation to Other Internet Scripting Languages

For many purposes, VBScript may simply be thought of as VBA for the Web. Certainly this appears to be the primary role initially envisioned for the language by Microsoft. It allows programmers versed in Visual Basic or VBA to leverage their expertise and experience to quickly create client-side and server-side Web applications. VBScript will run on Windows NT 4 and later, Windows 95 and later, Windows 2000, and Windows XP. However, because JavaScript had a head start as a client-side Web programming language, because native support for VBScript applications has not yet been added to the Netscape Web browser (while on the other hand, Microsoft Internet Explorer does natively support the use of JavaScript for client-side scripting), and because VBScript imposes generally fewer "sandbox" (i.e., security) restrictions than JavaScript, VBScript remains relatively less popular than JavaScript for Web client-side implementations. Because VBScript is not Web browser neutral, it is only useful for client-side scripting in organization Intranet environments where steps can be taken to ensure that Microsoft Internet Explorer is the browser used by those accessing the site. When used in such a scenario, experienced Visual Basic and VBA developers have found that VBScript does facilitate rapid development and implementation of powerful client-side scripts. While not a hit for client-side scripting, VBScript has proven extremely popular for server-side Web applications built on top of Microsoft IIS. ASP has shown itself flexible, robust, and better performing than nonscript approaches, and when used in conjunction with ASP, VBScript competes well with other scripting languages and scripting approaches.

The introduction of Windows Scripting Host in 1998 created a second important role for VBScript as the scripting language of the Windows operating system. Prior to 1998 many DOS and UNIX programmers had bemoaned the lack in Microsoft Windows of something truly equivalent in power and scope to a full-featured batch mode language or shell script language. Windows shortcuts and PIF files did not provide sufficient in-depth access to the operating system and its application programming interfaces. With the introduction of the WSH, VBScript now fills that void. VBScript allows programmers to write scripts for browsing portions of the Windows file system, bridging between Windows applications, and interacting with network resources such as database servers. Unfortunately such capability has come with a price. Many of the most infamous e-mail computer virus attacks involve the transmission and activation of VBScript files (file extension .vbs). While Microsoft has added automatic alerts and checks to warn users of potential problems and attacks, inexperienced and unsophisticated users still find themselves falling victim to such viruses all too frequently. However, in the proper hands and for the proper reasons, VBScript applications written to run in the context of the WSH can be most useful, particularly to accomplish quickly certain kinds of lightweight tasks involving the operating system and associated services.

For client-side Web applications JavaScript (or JScript as implemented by Microsoft) is clearly the best alternative to VBScript. For most common client-side tasks, e.g., Web form user input validation and manipulation, user warnings and alerts, soliciting user input through simple dialogs, and controlling Web browser windows, both VBScript and JavaScript provide more than adequate support. Both languages rely heavily on the World Wide Web Consortium (W3C) document object model (DOM). Both languages provide extensive support for Web browser events (e.g., page load and form submit). VBScript distinguishes itself from JavaScript mostly in terms of greater functionality and ease of use when integrating external programming classes and objects and accessing advanced client-side operating system features and functions. When such high-level functionality is required (and assuming there is assurance that the target audience will be using Microsoft Internet Explorer), VBScript is the better option.

For creating server-side Web applications using Microsoft IIS, there exist a number of alternatives to VBScript. Which one to use is largely a matter of personal preference. For example, IIS natively supports the use of the JScript language in ASP applications. Microsoft has committed to ensuring equivalent server-side script capability whether using VBScript or JScript. After download and installation of appropriate third party software (i.e., scripting engine), PerlScript or other languages also may be used instead of VBScript to code ASP applications. There also are numerous alternatives to ASP itself, including traditional Web common gateway interface (CGI) scripts (slower performance and no built-in support for maintaining state), applications that directly access Microsoft's Internet service application programming interface, and competing scripting-based schemes now compatible with IIS such as PHP tools and the Macromedia's commercial ColdFusion MX product.

Table 1 VBScript Major Version Release History

VBScript Version	Year of Release	Comment
1.0	1996	VBScript introduced in conjunction with version 3 release of Microsoft Internet Explorer.
2.0	1997	Coincided with version 3 release of IIS. Added several language enhancements, including support for instantiation and exploitation of external COM objects.
3.0	1998	Coincided with first release of WSH. No new language features, but the scripting runtime library for this release did add FileSystemObject.
4.0	1998	Coincided with release 6.0 of Microsoft Visual Studio. No notable new features were added with this release.
5.0	1999	Added intrinsic scripting engine support for regular expressions, added more features for object handling, and enabled the creation of object classes using VBScript.

How Microsoft Implements Scripting Languages

The Microsoft model of scripting distinguishes between "scripting hosts" and "scripting engines." Scripting engines process and execute scripts fed to them by scripting hosts. Scripting hosts are applications that provide script context and environment (including a host application object model) and instantiate instances of appropriate scripting engines to execute scripts. This modular approach to scripting allows multiple applications to use the same scripting engine. It also allows multiple scripting engines, each designed to handle a particular language of script, to support a single application (scripting host). Microsoft has supplied scripting engines for JScript and VBScript as part of the Windows operating system since release of Microsoft Windows 98. Scripting engines for PerlScript, REXX, and Python are available from third parties. Microsoft also provides sufficient information about scripting interfaces to allow programmers to build their own scripting engines (and scripting hosts).

While the scripting engine is responsible for actually running your script, it is the responsibility of the scripting host to provide the object model that can be used by your script to access features, functions, and services of the host application. At the present time Microsoft provides scripting hosts for Internet Explorer, IIS, Outlook, and the Windows operating system itself. Each of these hosts provides an object model specific to its application. Thus the Internet Explorer scripting host's object model is largely based around the W3C document object model and is used to script client-side features and functions of the Internet Explorer Web browser. The IIS scripting host's object model is used to read HTTP requests and generate HTTP responses. The WSH's object model provides access to essential operating system objects such as the environment, network, and standard input, output, and error devices. The Outlook scripting host's object model provides access to such Outlook objects as forms, editors, and converters. The object models for the Internet Explorer scripting hosts and the WSH are briefly described in this chapter. The object model for using VBScript in IIS is dealt with fully in the chapter on ASP and so is not dealt with here. The use of VBScript in Microsoft Outlook is outside the intended scope of this encyclopedia and so is also not addressed here.

In addition to object references provided by the scripting host, scripts written in VBScript have access to other properly registered Windows ActiveX and COM object classes on the host computer. This includes also objects intrinsic to the VBScript scripting engine (e.g., an object for processing regular expressions and an object for handling and processing runtime errors) and objects provided by the Microsoft Scripting Runtime Library (e.g., the Dictionary Object and the File System Object).

VBScript Version History

By design most VBScript upgrades have been synchronized with scripting host releases and upgrades (Microsoft, 2002b). This chapter describes VBScript through version 5.6, which was released in 2001. Effective with this release, version numbering for the WSH, VBScript, JScript, etc. have been made consistent. (Thus, VBScript release 5.6 coincides with WSH release 5.6.) The first three major version releases of VBScript coincided with first releases of new scripting hosts. Version 4 was released to coincide with the release of version 6.0 of Microsoft Visual Studio, but added no notable new language features. Version 5 added several new language features and has been maintained for more than two years with minor release updates, most associated with updates in scripting hosts. With the release of Visual Basic.NET, there is no longer a clear need for a separate Visual Basic scripting language. Release 5.6 may be the last release of VBScript. Table 1 details major version release history of VBScript.

VBSCRIPT LANGUAGE FUNDAMENTALS

VBScript Variables, Constants, and Data Typing

Variables and named constants are constructs used to represent data employed in a program. Each variable or named constant is known in code by a symbolic name. Variables differ from named constants in that variable values may change during program execution and initial variable values may be the result of calculations or manipulations carried on during program execution. Constants are declared in VBScript using the key word

"Const," remain fixed throughout script execution, and are typically initialized in a declaration statement, e.g.,

```
Const MyConst = "This is an unchanging
  string."
```

VBScript variables are declared using the any of the key words "Dim," "Public," or "Private." (The key words Public and Private are most often chosen to declare variables in script modules that are to be used as objects by other programs or scripts as discussed later.) VBScript supports implicit declaration of scalar variables; i.e., it does not require that scalar variables be declared before first use. Both fixed array variables and dynamic array variables (i.e., arrays whose sizes may be changed during program execution) are supported in VBScript; however, unlike Visual Basic and VBA, the base value of all array indices is always zero and cannot be changed. Care must also be taken in VBScript when assigning arrays as values of items in an object collection. Object collections are themselves implemented in a fashion reminiscent of arrays, resulting in syntactical ambiguities when one tries to reference a specific element of an array that is itself an item in an object collection. (The solution is to assign in its entirety the collection item that is an array to another, stand-alone array variable, and then to access individual array items through that stand-alone array variable. The stand-alone array variable can be reassigned to the collection item variable after any value changes.) Unlike Visual Basic and VBA, VBScript does not support user-defined structured data types. VBScript variable and constant names are not case sensitive and must not be reserved words. VBScript includes a number of intrinsic constants (e.g., vbCrLf, which is the standard PC end-of-line 2-byte string) all of which begin with the letter sequence "vb." These constants cannot be reassigned nor may their names be used as variable names (e.g., the statement vbCrLf = "hello" will generate a run-time error). Intrinsic constants are provided for output display color values (e.g., vbBlack), comparison operations (e.g., vbTextCompare to do a case-insensitive text comparison), date values (e.g., vbSunday), and string and character values (e.g., vbTab). Programmers may also import constant declarations (e.g., when working with supplemental or third-party software libraries). Finally, VBScript variables and named constants do have scope; i.e., variables declared within a function or subprocedure are local in scope and are not accessible to code outside that function or subprocedure. In applications where script modules are to be used as objects by other programs or scripts, variables also have public (variable may be accessed by applications invoking the script module as an object) or private (variable is only available within the script module in which it is declared) scope.

Because computers can store a variety of data in a variety of different formats and representations, each variable or named constant typically has associated with it a data type. The data type defines the kind of data (e.g., number, character, logical, and object reference) and often its structure (e.g., 2-byte integer, multicharacter string, and integer array). Unlike Visual Basic and VBA, however, VBScript recognizes only one primary data type, which is called a variant. Variant is a special chameleon-like data type that can be used for any type of data, even including objects. A variant is actually made up of two components, one containing the data itself (or a pointer to the data, e.g., in the case of objects, multibyte numbers, strings, and arrays) and a second part, called the variant subtype, indicating the representation of the data stored, i.e., its effective data type. A variant can be used at one point in a procedure to store an integer number and then used later in the same procedure to store a string (though doing this is usually bad practice). Table 2 lists available variant subtypes.

Variants are useful shortcuts when scripting because they let the scripting engine worry about the optimal representation in the storage of an item of data. However, because variants encourage programmers to ignore variable data types when authoring code, it is easy to inadvertently

Table 2 VBScript Variant Subtypes

Variant Subtype	Comment
Empty	Variable has been declared but has not yet been initialized.
Null	Contains no data (as result of an explicit assignment or assignment to result of an expression involving other null variables).
Boolean	Contains True or False.
Byte	Contains 8-bit value (i.e., between 0 and 255).
Integer	Signed whole number in 2 bytes (i.e., between –32,768 and 32,767).
Long	Signed whole number in 4 bytes (i.e., between –2,147,483,648 and 2,147,483, 647).
Single	Signed, single precision floating point number (i.e., can be used to express nonwhole real numbers) expressed in 4 bytes.
Double	Signed, double precision floating point number (8 bytes).
Currency	Signed, fixed decimal number (8 bytes) used for storing currency values.
Date / Time	Contains date and time between January 1, 100 and December 31, 9999.
String	Variable length string of characters (up to 2 billion characters in length).
Object	Contains reference to an object.
Array	Contains reference to an array.
Error	Contains error number (such as generated by scripting engine).

generate data type mismatch errors when doing comparison operations and operations involving multiple operands or arguments. To avoid uncertainty, it is possible to ascertain the variant subtype of a variable and even to force transformation of data stored in a variable from one variant subtype to another. Variant (and named constant) subtype can be determined by using either the TypeName function or the VarType function (the former returns a human-readable string indicating the subtype, the latter a number code for the subtype). Values of variants can be transformed (cast) using a variety of type-specific functions, most of which begin with the letter "C" (e.g., CDate, CStr, and CInt). Listing 1 gives the code for a simple VBScript file that can be run under WSH. This script will generate three pairs of message boxes. The first message box in each pair shows that the value of the variable "myVar" is the numeral or number one. The second message box in each pair shows the variant subtype of myVar at each point in the program. (To run this script yourself, create a plain text file on your computer containing the code as shown and save it with the .vbs extension. Then double click on the file name from Windows Explorer or enter the file name at the command prompt.)

Listing 1: VBScript illustrating use of TypeName and CStr functions.

```
Option Explicit
Dim myVar

REM myVar as variant subtype String
myVar = "1"
MsgBox (myVar)
MsgBox (TypeName(myVar))

REM myVar as variant subtype Intege
myVar = 1
MsgBox (myVar)
MsgBox (TypeName(myVar))

REM myVar transformed from subtype Integer
  to subtype String
myVar = CStr(myVar)
MsgBox (myVar)
MsgBox (TypeName(myVar))
```

Essential Language Elements

VBScript statements are line delineated, i.e., one statement per line unless special line continuation or statement separator characters are used. For this reason no end of statement character delimiter is used in VBScript.

VBScript provides a typical set of arithmetic operators for addition, subtraction, division, and multiplication (+, −, /, ∗), as well as special arithmetic operators for no remainder integer division (\), remainder only modulo division (Mod), and exponentiation (^). The string concatenation operator is "&" (though where unambiguous "+" can be used instead). Comparison operators are >, <, < >, >= , < = , and = . Comparison operators can be used to compare strings as well as numbers. "And," "Or," "Not," and "Xor" (exclusive Or) are used both as Boolean logical operators and to perform bitwise operations. In the absence of parentheses, operators are evaluated left to right in the following order: arithmetic (exponentiation, division and multiplication, integer division, modulo arithmetic, and addition and subtraction), concatenation, comparison, and finally logical/bitwise operations (Not, And, Or, and Xor).

VBScript version 5.6 provides a rich set of program flow control constructs. Branching can be accomplished using block-style If . . . Then . . . Else and Select Case statements. Listing 2 shows a simple VBScript file that solicits an integer between 1 and 3 from the user and branches accordingly using an If . . . Then . . . Else statement. Listing 3 does the same job using a Select Case statement. VBScript version 5.6 also supports flow control through the use of several kinds of looping structures. VBScript Do . . . Loop statements allow for either top or bottom tested loops using either "until" or "while" conditionals. Thus "Do While intMyValue > 0 . . . Loop" will execute the included block of statements repetitively until the value contained in the variable intMyValue becomes zero or less than zero. The statement block will not be executed at all if intMyValue is zero or negative when the Do statement is first encountered. Conversely the block of statements contained within "Do . . . Loop Until intMyValue > 0" will always be executed at least once and will be executed repetitively until the value of intMyValue is positive. Errors will arise if the variant subtype of intMyValue is or changes to be nonnumeric. Also, unlike Visual Basic, VBScript does not support the "Do Events" statement, which means that once inside a loop, scripts do not respond to external events other than those checked explicitly within the loop. If a coding error results in an endless loop this means that script execution can only be terminated by termination of the scripting host application itself. Listings 2 and 3 illustrate a simple use of a "Do . . . Loop" construct. VBScript also supports for . . . next and for each . . . next loops. These types of loops are used to execute blocks of statements a fixed number of times (for . . . next) or until all the values in a given collection have been processed (for each . . . next).

VBScript allows limited use of Visual Basic's On Error statement to implement user-defined error handling. Specifically, VBScript allows "On Error Resume Next" and "On Error GoTo 0" forms of this statement. (No other uses of On Error are supported in VBScript.) Most errors in script statements that occur after an "On Error Resume Next" statement and before the next "On Error GoTo 0" statement will be left for the programmer to handle and will not result in program termination. Information about any errors that occur is available to the programmer via the "Err" object (discussed later).

Finally VBScript includes a number of intrinsic functions for manipulating data and interacting with the host system. These are too numerous to describe in detail here, but they include functions for obtaining system date and time information, transforming variant subtypes, determining whether the content of a variable is or can be meaningfully transformed to a particular variant subtype, manipulating and calculating mathematical quantities, manipulating date and time information (e.g., adding to and subtracting from dates), and manipulating string and character data. Listings 2 and 3 illustrate the use of the

Figure 1: What the user sees when running either of the scripts given in Listings 2 and 3.

IsNumeric() function which returns a Boolean value indicating if data are or can be transformed into a numeric variant subtype. Note that what the user sees is identical regardless of which of these two scripts is executed. Figure 1 shows what the user sees when running either of the scripts given in Listings 2 and 3.

Listing 2: VBScript illustrating If... Then... Else program flow control.

```
Option Explicit
Dim strMyVar, intMyVar

Do
    strMyVar = InputBox_
      ("Enter a number between 1 and 3")
Loop Until IsNumeric (strMyVar)

intMyVar = CInt(strMyVar)
If intMyVar = 1 then
  MsgBox ("You Entered the Lowest Option.")
ElseIf intMyVar = 2 then
  MsgBox ("You Entered the Middle Option.")
ElseIf intMyVar = 3 then
  MsgBox ("You Entered the Highest Option.")
Else
  MsgBox_
    ("You Entered a Number Too Small or Too
    Big!")
End If
```

Listing 3: Script illustrating Select Case program flow control.

```
Option Explicit
Dim strMyVar, intMyVar

Do
    strMyVar = InputBox_
      ("Enter a number between 1 and 3")
Loop Until IsNumeric (strMyVar)
intMyVar = CInt(strMyVar)

Select Case intMyVar
Case 1
  MsgBox ("You Entered the Lowest Option.")
Case 2
  MsgBox ("You Entered the Middle Option.")
Case 3
  MsgBox ("You Entered the Highest Option.")
Case Else
  MsgBox_
    ("You Entered a Number Too Small or Too
    Big!")
End Select
```

User-Defined Procedures

In addition to intrinsic functions, VBScript allows programmers to define their own functions and subroutine procedures. User-defined functions, code blocks that begin with the key word "Function" and end with an "End Function" statement, return a single value. Since VBScript only allows variant data type, all functions return a variant; therefore a data type is not included in a function declaration (as it typically is when declaring functions in Visual Basic and VBA). User-defined subroutines, code blocks that begin with the key word "Sub" and end with an "End Sub" statement, do not return a specific value, but rather perform tasks (including input, output, and user dialog tasks) and may affect values of script-level (i.e., global) scope variables. Variables and named constants declared in functions and subroutines have a procedure-level (i.e., local) scope and are not directly visible to the calling routine. (However care must be taken not to try to use a script-level variable with the same name as a variable local to a particular subroutine or function.) Local scope variable values set or changed during one call to a subroutine or function do not carry over to subsequent calls. Subroutines and functions may be called recursively (i.e., may call themselves), though great care must be exercised when doing so. By default, argument variables needed by a user-defined subroutine or function are passed by reference. This means a pointer to each argument variable in memory is passed. Any modifications to an argument variable that happen in a subroutine or function persist when control is passed back to the calling procedure. A programmer may use the ByVal key word in the subroutine or function declaration—e.g., "Sub MySub (ByVal myVar)"—in order to override this default behavior. Arguments passed by value remain unchanged when control passes back to the calling procedure, regardless of what is done in the subprocedure.

VBScript Coding Conventions

Over time, the community of programmers using any given language tends to develop conventions in areas such as naming patterns for variables and constants, the way program commentary is incorporated, and the way code is formatted and indented. While there is no compulsion to

follow such conventions, doing so tends to improve readability and sharability of code. There are several recognized coding conventions used by VBScript programmers (Microsoft, 2002c). Two naming patterns are commonly used for named constants. One pattern specifies all upper case for names of constants. When using this pattern, individual words within the name of a constant are delineated using an underscore (thus, "MY_CONSTANT_NAME"). An alternative approach is to prefix names of constants with "con" and use a mixed case pattern where the first letter of embedded words is upper case (thus, "conMyConstantName"). This latter approach follows from variable naming conventions, where a three-letter prefix is used to indicate the variant subtype of the variable (i.e., strMyStringVariable). Sometimes an additional character ("s") is prefixed to names of variables having script-level scope (thus, sintMyIntegerVariable). User-defined subroutine and function names are also typically mixed case and usually begin with a verb describing what the procedure does—e.g., "Sub FetchEnvironmentValues ()".

Lines prefaced by either the string "REM" or a single apostrophe are interpreted in VBScript as in-program comments and are not processed by the VBScript interpreter. A single apostrophe following a VBScript statement is also interpreted as the start of a comment. To use the REM key word on the same line following a VBScript statement, the REM key word must be preceded by a colon. (A colon is the delimiter used to allow multiple VBScript statements on a single line.) All text following either the key word REM or a single apostrophe is ignored by the scripting engine to the end of the current line. In-program comments not only should be descriptive, but also should focus on what rather than how, in order to minimize comment maintenance. Thus each procedure or subprocedure should begin with an extended comment describing the purpose of the procedure, its required input(s), any outputs, any assumptions about inputs and outputs, and its return and/or the global variables effected. A minimum should be said about how it accomplishes these tasks since that information is typically more volatile.

When scripts are created, indenting should be used to show flow control and logical structure of the code. Thus blocks of code within an If . . . Then . . . Else statement (and similar flow control statements) will be indented relative to the If and Else lines themselves. To conserve screen space and allow for nesting of code blocks, a typical indent is only three to five spaces.

Finally, though not required by the language, many programmers find it useful to declare (using Dim statements) all variables and named constants before use. Starting a script off with the line "Option Explicit" requires that all variables and named constants used in that script be declared before first use. If a typo in a name is made in the code, having option explicit in effect will immediately highlight the error, which can greatly facilitate code debugging.

OBJECTS IN VBSCRIPT

Use of Objects in VBScript

Objects are abstract structures used in high-level programming languages to encapsulate discrete modules of code and associated data structures. They enable modular implementation and reuse of code and facilitate collaborative development of advanced applications. When the archetype of a programming object, called an object class, is written, standard types of programming interfaces and class behaviors are defined that allow instances of objects created from the class definition to be used easily and reliably in many different programs. Instead of writing their own code to perform common tasks, programmers can use classes as templates to instantiate (create) object instances for use in their programs. Programmers implementing a predefined class can treat an object instance created from that class like a "black box." A programmer using an object class does not have to understand how the class works internally. All he or she has to understand is how to use the interfaces and behaviors of the class. Interfaces and behaviors of external object classes invoked using VBScript are described in terms of properties, methods, and events. Object classes written in more advanced object-oriented programming languages (e.g., C++) may also exhibit polymorphism and inheritance. These latter two behaviors are not supported in VBScript; however, they are supported in Visual Basic.NET.

VBScript provides support for the use of registered ActiveX object classes, also referred to as object linking and embedding server objects and COM or COM+ objects. While VBScript provides a language intrinsic CreateObject method for instantiating a new instance of an object and a language intrinsic GetObject method for creating additional references to an already instantiated object, it is usually better when possible to invoke equivalent methods built into an application-specific object model provided by the scripting host being used. This allows the scripting host to track the use of the object and resources consumed by it and integrate the object into the host application environment. Thus, when a script is written for the WSH, the preferred syntax to create an instance of an Active Data Object (ADO) Recordset object would be

```
Set objMyRecordset = WScript.CreateObject
  ("ADODB.Recordset")
```

Note that the string argument of the WScript object's CreateObject method is the ProgID of the object, typically consisting of the library or ActiveX server name and the type name (class name) of the object being instantiated. ProgIDs are maintained in the HKEY_CLASSES_ROOT key of the host system's registry. Note the use of the "Set" key word in the statement that assigns the reference for the object to the variable objMyRecordset. This is how VBScript knows you are assigning an object reference to a variable. When finished with an object, you should use the "Nothing" key word to free resources assigned to the variable; thus,

```
Set objMyRecordset = Nothing
```

Note also that VBScript uses dot syntax for referencing an object's methods and properties (including collections of properties) and underscores syntax for referencing an object's events. Thus in this example WScript.CreateObject() referenced the CreateObject

method of the WSH object model's WScript object. Reference to WScript.ScriptFullName returns the variant (subtype string) property of the WScript object that contains the full path and file name of the script being executed. WScript.Arguments.Item(0) returns the first item in the list of command line arguments used when the script was invoked (an example of a collection). Object event references are typically used in user-defined subroutine declarations. Thus, on a Web page containing an HTML <Select> element (i.e., pull-down menu) named "myChoice," a programmer could include a subroutine declared as "Sub myChoice_OnChange (),"which would then be invoked each time the user changed his or her selection from that particular menu.

Intrinsic Object Models

VBScript provides two built-in object models, one intrinsic to the language itself, and one included as part of the Microsoft Scripting Runtime library (and therefore also available to JScript implementations). The former includes the regular expression object (RegExp) and the error object (Err). Note that the RegExp object must be instantiated using the New key word (as typical for early binding) rather than the CreateObject method (late binding) as required for other objects instantiated in VBScript; thus,

```
set objMyRegularExpression = New RegExp
```

The RegExp object allows testing of strings for pattern matches and manipulation of matches found through the "Match" object collection (a child of the RegExp object class).

The Err object is singular, is always available, and therefore does not need to be explicitly instantiated. The Err object includes a method to clear an error condition and properties providing the number and description of the last error that occurred. It is used in conjunction with the On Error Resume Next statement for user-defined error handling.

Version 5.6 of the Microsoft Scripting Runtime Library provides two important objects, the FileSystemObject, for accessing file system directory information and reading from and writing to individual text files, and the Dictionary object, which is similar to associative arrays found in Perl and other programming languages. The FileSystemObject is at the top of an extensive and powerful object hierarchy that provides access through a variety of child objects to the host system's drives, directories, and files. "Scripting.FileSystemObject" is the ProgID for the FileSystemObject. Using methods and properties of the FileSystemObject you can browse drives and folders of the host computer, create new folders and text files on the host computer's file system, and open existing text files on the host computer for reading and/or writing (e.g., by creating and using child TextStream objects). Note, when using the FileSystemObject (and most other external COM objects) for client-side scripting in Microsoft Internet Explorer, Internet Explorer typically warns users that use of these objects might be unsafe. Users can even configure Internet Explorer not to allow instantiation and use of external COM objects. Listing 4 illustrates the use of the FileSystemObject in WSH.

Finally, the Dictionary object of the Microsoft Scripting Runtime Library allows you to store and manipulate lists of key value pairs, i.e., lists of values where each value is tied to a unique key. "Scripting.Dictionary" is the ProgID of the Dictionary object. This object has a small set of properties and methods that allow your script to add and remove items from the dictionary list, retrieve or alter individual values by key, change or replace individual keys, and check to see if a particular key exists in the dictionary list.

Listing 4: Illustrating the use of FileSystemObject in the context of WSH.

```
Option Explicit
Const conMyLine =_
  "This was written by a VBScript script."
Const conMyFileName = "Temp.txt"
Dim objMyFS, objMyTS
Dim strFileName, strTextLine
set objMyFS = WScript.CreateObject_
  ("Scripting.FileSystemObject")
strFileName = Replace_
  (WScript.ScriptFull Name,_
  WScript.ScriptName, conMyFileName)

if objMyFS.FileExists(strFileName) then
    set objMyTS = objMyFS.OpenTextFile_
      (strFileName)
    strTextLine = objMyTS.ReadLine
    objMyTS.Close
    MsgBox (strTextLine)
else
    set objMyTS = objMyFS.CreateTextFile_
      (strFileName)
    objMyTS.WriteLine(conMyLine)
    objMyTS.Close
    MsgBox ("Created" & strFileName)
end if

set objMyTS = Nothing
set objMyFS = Nothing
```

Windows Scripting Host Object Model

The WSH object model consists of a top-level object, WScript, with properties germane to the script and how it was invoked, methods for suspending and ending script execution and creating and using external objects, and a half-dozen immediate child objects, each providing access to a service or a function of the host system. WScript child objects include: WshArguments, a collection containing any command line arguments used when invoking the script; WshNetwork, providing access to network configuration and services; StdIn, StdOut, and StdErr, specialized read-only or write-only (as appropriate) TextStream objects providing access to standard input and output devices (e.g., keyboard and display); and WshShell, providing access to a variety of Windows

shell services including the registry, log files, environment variable values, and system folders (the latter two functions are implemented through the use of child objects, WshEnvironment, and WshSpecialFolders). The WshShell object also allows your script to create and save shortcuts, invoke other applications, and simulate the sending of keystrokes to another application. Listing 4 provides the code for a script that uses the WScript.ScriptFullName and WScript.ScriptName properties and the intrinsic VBScript "Replace()" function to construct a reference to a file named "temp.txt" located in the same directory as the script. (This assumes that the value of WScript.ScriptName appears exactly once in value of WScript.ScriptFullName.) Using the Scripting Runtime Library's FileSystemObject this script then displays the first line of text contained in temp.txt, or if the file does not exist, it creates the file, writes text to it, and then saves it.

The spate of computer viruses that exploited VBScript and the WSH, led Microsoft to introduce a new, beefed-up security model as part of WSH 5.6 (Clinick, 2000). This new model allows users (or system administrators) to set a trust policy for scripts submitted to WSH. Using a particular registry key (\HKEY_CURRENT_USER\SOFTWARE\Microsoft\Windows Script Host\Settings\TrustPolicy), users can enable running of any script in WSH, allow only scripts signed by a trusted third party (i.e., someone in your Trusted Publishers List) to run under WSH on your machine, or prompt before a script not signed by a trusted third party is run. To sign scripts you must have a valid certificate (typically from a commercial certification authority). Script signing is accomplished programmatically using the Signer object, which is part of the latest Microsoft Scripting Runtime Library.

Internet Explorer Scripting Host Object Model

The VBScript scripting host object model for Internet Explorer differs significantly from those for IIS, WSH, and Outlook. As just mentioned it is only implemented on Internet Explorer, and it is not presently available on Netscape Navigator. Client-side script is embedded within HTML (specifically inside the HTML <Script> element), and it is intended to interact closely with HTML objects (elements). The Internet Explorer scripting host object model is derived largely from the W3C DOM for HTML, but it adds more support for events and a few other features to facilitate use of the DOM with scripting. Object events play a particularly important role in the Internet Explorer object model. The top-level object in the Internet Explorer scripting host object model is the "Window" object (referring to the Web browser window, not the operating system). The most important child of Window is the "Document" object, which contains collections and child objects relating to all the HTML elements on the Web page displayed by the browser, including especially all objects in any HTML form or forms on the page (i.e., those with which the user can interact). The Window object also includes properties and child objects that provide context for the document (e.g., current URL, previously viewed pages, information about frames in use, parent windows, etc.).

Client-side VBScript can be used for a variety of purposes, but most commonly it is used for validation of data entry on HTML forms, manipulation of HTML content on a page (e.g., hiding and showing certain options, changing text and values of elements and/or form menus in response to user actions), and extended interactions with Web clients beyond those inherently supported by the Web browser itself (including those invoked from external COM objects). For example, VBScript can be used to validate input data at time of form submittal, to validate each entry and menu selection as it is being made, to update menus and form structure as specific selections are made (e.g., to give only compatible selections), or to facilitate browsing of a resource via a table of contents. An illustration of the last scenario would be an application that provided two frames, one smaller, narrower frame showing the table of contents for the work, and a second larger frame for showing the text of a given subsection of the work. Script could then be written to allow the user to expand and hide sections of the table of contents and to synchronize the text shown in the larger frame as items in the table of contents frame are selected and deselected.

The complete Internet Explorer object model is too extensive to document in its entirety here, but a few specifics are worth highlighting. Script not contained in a subroutine (e.g., not intended to handle a specific event) is executed as encountered and can be used to help construct or modify the HTML page as it is being rendered on the client side. Other script is invoked only in response to certain events, e.g., page loading or unloading, form submittal, or events associated with form elements such as when a text box gets focus, when a check box is checked, when a selection from a menu is made or changed. The same HTML element may be examined and/or manipulated using different components of the Internet Explorer object model. (Thus a form element may be accessed either through the Window.Document.All collection or through the Window.Document.Forms collection.) Available properties, methods, and events may be different for different HTML elements, though nearly all HTML elements share properties such as InnerHTML (content of the element including HTML markup), InnerText (content of the element stripped of HTML markup), ParentElement (immediate HTML parent to the element), NextSibling (next HTML element having same parent), and PreviousSibling (previous HTML element having same parent), and collections such as Attributes (collection of all attributes of the element), ChildNodes (collection of all HTML elements that are immediate children of the element), and All (collection of all HTML elements contained in the element). Generally, to be most easily manipulated in script, HTML elements should be assigned unique id attributes (all elements on an HTML page) and/or unique name attributes (form and anchor elements). HTML form elements typically will have values and text content associated which may change in response to user actions. For instance, checkable form elements (e.g., radio buttons and check boxes) will have a Boolean checked property to indicate whether the button or box has been checked, while

select menus will have a selected index property to indicate which menu item(s) has been selected. Script can be used to insert HTML elements into the DOM for a page at the time the page is loaded or later, in response to a particular user action.

Selection of which components of the Internet Explorer object model to use depends largely on the purpose of the script, the exact nature of the HTML elements to be examined or manipulated, and the way the HTML is written. For instance, the Window.Document.All collection provides access to every HTML element on the Web page. Using the All collection, HTML elements may be accessed by value of id attribute (if present), by the element's position on the page (i.e., the sequence of the elements on the page relative to all HTML elements on the page), by value of the name attribute (but only if the element is contained in a form), or by element tag name. When retrieving by name attribute, tag name, or id attribute (since for HTML Internet Explorer does not enforce the uniqueness of id attributes), a subcollection may be retrieved rather than a single HTML element reference if more than one element has the tag name or attribute specified. To get the specific HTML element of interest requires that the programmer either also specify its sequence among all those with the same name attribute, tag name, or id attribute or iterate through the subcollection until the HTML element with the desired property is found.

Among alternatives to the Window.Document.All collection are: the anchors collection, which gives access to all HTML <a> elements having nonnull name attribute value and/or nonnull id attribute value; the links collection, all HTML <a> elements having a nonnull href attribute, and all HTML <area> elements; the forms collection, all HTML <form> elements (including children); the images collection, all HTML <img> elements; and several other specialized subcollections of HTML page elements. Note also that the Internet Explorer Object model supports shorthand syntax. For instance, forms can be referred to by name directly or as an item in the Window.Document.Forms collection. Thus, the HTML form element "<Form name = 'myForm'>" may be referenced in script using either "Window.Document.myForm" or Window.Document.Forms.Item("myForm")." Other forms of shorthand are supported as well, though overuse of such shorthand can reduce code clarity. Listing 5 illustrates a client-side VBScript that uses script to populate a select menu on a form when the page is first loaded. VBScript is also used to insert an additional label and text box on the form when "other" is selected from the pull-down menu by the user. Figure 2 shows what the user sees when running the script in Listing 5.

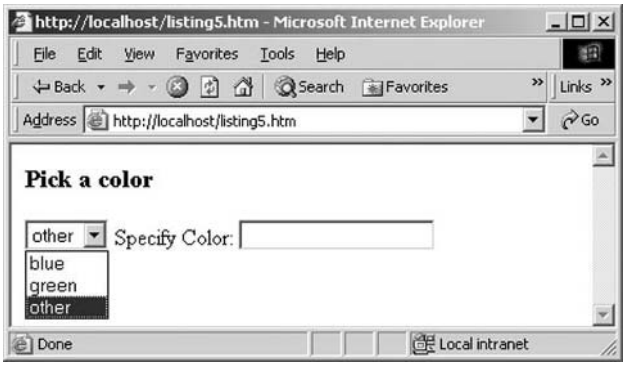


Figure 2: What the user sees when running the script in Listing 5.

Listing 5: Illustrating client-side VBScript in Internet Explorer.

```
<html>
<head><script Language='VBScript'>
sub window_onLoad
   set objMyTag = Window.Document.CreateElement("option")
   Window.Document.myForm.Item("myMenu").Options.Add (objMyTag)
   objMyTag.innerText="blue"
   set objMyTag = Window.Document.CreateElement("option")
   Window.Document.myForm.Item("myMenu").Options.Add (objMyTag)
   objMyTag.innerText="green"
   set objMyTag = Window.Document.CreateElement("option")
   Window.Document.myForm.Item("myMenu").Options.Add (objMyTag)
   objMyTag.innerText="other"
end sub
sub myMenu_onChange
   if Window.Document.myForm.Item("myMenu").SelectedIndex = 2 then
     set objMyTag = Window.Document.CreateElement("span")
     document.forms.item("myForm").appendChild (objMyTag)
     objMyTag.innerText = "Specify Color:"
     set objMyAtt = Window.Document.CreateAttribute("id")
     objMyAtt.value = "OtherColorLabel"
     objMyTag.Attributes.setNamedItem (objMyAtt)
     set objMyTag = Window.Document.CreateElement("Input")
     document.forms.item("myForm").appendChild (objMyTag)
     set objMyAtt = Window.Document.CreateAttribute("type")
```

```
    objMyAtt.value = "text"
    objMyTag.Attributes.setNamedItem (objMyAtt)
    set objMyAtt = Window.Document.CreateAttribute("name")
    objMyAtt.value = "otherColor"
    objMyTag.Attributes.setNamedItem (objMyAtt)
    set objMyAtt = Window.Document.CreateAttribute("id")
    objMyAtt.value = "OtherColorTextbox"
    objMyTag.Attributes.setNamedItem (objMyAtt)
  else
    if Window.Document.myForm.childNodes.length = 5 then
       Window.Document.All.Item("OtherColorLabel").RemoveNode(True)
       Window.Document.All.Item("OtherColorTextbox").RemoveNode(True)
    end if
  end if
end sub
</script></head>
<body>
<form name='myForm' action=" method='Get'>
  <h3>Pick a color</h3>
  <select id='myMenu' name='color'></select>
</form></body></html>
```

ADVANCED FEATURES OF VBSCRIPT

Object classes enable encapsulation of programming code, which in turn facilitates modular development of applications and the reuse of code. While VBScript has long provided stable support for the use of object classes by script (i.e., since VBScript version 2), mechanisms to support modular development and reuse of script itself were less stable and less robust prior to the release of VBScript version 5. VBScript now provides several ways of varying complexity and sophistication to modularize script development and deployment and facilitate script reuse. Discussed briefly in this section are three specific technologies that facilitate script modularization and reuse: Windows Script Files, Remote Scripting, and Windows Script Components. In addition, VBScript, as of version 5, supports class declaration statements, and these statements allow for the direct creation of full-blown object classes written entirely in VBScript. Class declarations are mentioned here in the context of remote scripting, but a more complete discussion of authoring generic classes using VBScript is beyond the scope of this chapter.

Windows Script Files

Windows Script Files (typical extension.wsf) are XML files that facilitate the simple modularization and reuse of scripts written for WSH. Windows Script Files contain a mix of script, contained in <Script> elements, and XML elements that define the flow and use of both internally contained script as well as script contained in external files (e.g., files with extensions such as .vbs, .pl, and .js). Windows Script Files allow script authors to incorporate libraries of previously written script functions and subprocedures into current projects and jobs. Windows Script Files support the development and deployment of individual script applications that use multiple scripting engines (i.e., are written in multiple scripting languages) and the import of type libraries associated with external objects and containing reusable constant declarations relevant to those objects. Windows Script Files can be created or edited using a standard XML or plain text editor, and single files may contain instructions for multiple script jobs or projects. Files with extension.wsf are automatically associated with the WSH. They can be invoked from either the command line or the Windows Run dialog. Command line switches are available that allow the user to run Windows Script Files in batch mode (i.e., suppresses error messages, MsgBox dialogs, InputBox dialogs, etc.), in debug mode, and with a time-out. Command line switches also allow a user to specify scripting engine to use and which job to run (for a multijob Windows Script File). Windows Script Files are further described, including examples, on the Microsoft MSDN Web site (Microsoft, 2002a) under Web Development, Scripting, Documentation, Windows Script Technologies, Windows Script Host, Running Your Scripts.

Remote Scripting

Remote scripting has been available to JScript programmers for some time, and most extant examples available today rely on JScript. However, the inclusion of the class declaration statement in VBScript version 5 means that this technology is now also available to VBScript programmers (Clinick, 1999). Remote scripting is an implementation by which client-side scripts running in the context of a Web browser (e.g., Internet Explorer) can call methods of objects written in server-side script (i.e., methods of ASP pages running in context of IIS), or if preferred, can actually instantiate ASP pages running on the server as objects in client-side script. For this to work, the VBScript on the ASP page must define a class and then communicate its availability to the Remote Scripting Runtime. It does so by creating an object named public_description, initializing that object as an instance

of the class to be exposed, and calling RSDispatch. A generic Java applet (Rsproxy.class) is used as a communication conduit to convey client-side script requests to the IIS Web server, which then invokes the ASP page containing the method being called and responds to the applet, which then passes the result of the call back to the client-side script. The client-side call can be made synchronously (i.e., locks the Web browser until response is received) or asynchronously (i.e., allows client-side script processing to continue, calling the application back when the result is ready). Examples and more details about how remote scripting works are available on the Microsoft MSDN Web site (Microsoft, 2002a) under Using Remote Scripting. It should be noted, however, that remote scripting is functionally a precursor to what is called in the .NET framework Web services. As such, remote scripting will almost certainly be supplanted eventually by technologies such as .NET Web services and the more generic simple object access protocol (SOAP).

Windows Script Components

Windows Script Components are a generic, object-based way to encapsulate script fragments and modules. The roots of Windows Script Components go back to earlier versions of VBScript, but in their present form they are supported as of VBScript version 5. Windows Script Components are implemented through the Microsoft Script Component Runtime Dynamic-Link Library (i.e., SCROBJ.DLL, which ships with latest version of VBScript). This library includes intrinsic object interface handlers suitable for use with IIS scripting host (i.e., ASP), Internet Explorer scripting host (i.e., Dynamic HTML), and standard COM Automation hosts (e.g., WSH) and is extensible. In essence, applications, including other scripts that want to make use of your Windows Script Components access your components through the Script Component Runtime Library. Your Windows Script Components themselves (a typical file extension is .wsc) are simply XML files that contain your script in combination with XML elements that describe your component, including which interface handler is to be used, to the script component run time. The top-level XML element of a Windows Script Component file is <package>, which in turn contains one or more <component> elements. As appropriate <component> elements may contain any of the elements listed in Table 3.

Script Encoding

Being able to modularize and reuse script has created a demand for ways to protect the integrity of script fragments and to help protect their potential value as intellectual property. Toward this end Microsoft has introduced the "Script Encoder" utility, a simple command line tool (SRCENC.EXE) downloadable from Microsoft that can be used to encode VBScript files for use with script intended for the WSH, the IIS scripting host (i.e., ASP), and the Internet Explorer scripting host. The Microsoft Scripting Runtime Library provides the utilities used by scripting hosts/scripting engines to decrypt the encoded script. The SCRENC utility is invoked with command line arguments specifying optional switches (e.g., whether to provide diagnostic output during encoding), the input file name (i.e., file to be encoded), and output file name (i.e., file containing encoded script). SCRENC will encode both VBScript and JScript sources. Language attribute values on any <script> elements containing encoded script and ASP page @Language directive values are automatically changed to "VBScript.Encode" or "JScript.Encode" as appropriate. Script authors must identify the fragments of script in the input source file that they want encoded by use of an embedded comment marker immediately before each segment of script to be encoded (e.g., **Start Encode**). Script encoding is further described, including examples, on the Microsoft MSDN Web site (Microsoft, 2002a) under Scripting, Windows Script Technologies, Script Encoder (Scripting Runtime Library).

MIGRATING FROM VBSCRIPT TO THE .NET FRAMEWORK

.NET Programming Language Environment

The Microsoft .NET framework represents a major reengineering of Microsoft's approach to support for application development and implementation. .NET introduces a new metalanguage called Microsoft intermediate language (MSIL), a just-in-time (JIT) compilation scheme, and a new common language run time (CLR) for the Windows platform. The .NET framework continues support for scripting (since beta version 2), but the fundamental nature of script implementation under .NET has changed such that it is no longer necessary to maintain VBScript as a separate language distinct from Visual Basic. Visual

Table 3 Allowed Child Elements of Component Element in WSC Files

Element	Comment
<registration>	Containing information like ProgID, as required for the COM Automation interface handler.
<public>	Describing publicly accessible properties, methods, and events associated with your component.
<implements>	Specifying script component runtime interface handler for your component.
<script>	Containing the VBScript that implements the logic of your component.
<object>	Creates an instance of an external object for use by your script.
<reference>	Links to an external type library.
<resource>	Containing declarations and initialization for values used in your script that you would rather initialize externally to the script itself.
<comment>	Ignored during component parsing and execution.

Basic programmers who want to script under .NET (e.g., in ASP.NET pages) use the same Visual Basic .NET language as when authoring stand-alone code meant to be compiled and run as a .EXE file. This adds considerable power for script authors to exploit, but it arguably does so at the cost of increased complexity.

The .NET framework itself, downloadable for current versions of Microsoft Windows as the .NET framework redistributable, is a set of extensible object classes providing comprehensive access to the events and services of the Microsoft Windows platform. Because it supports multiple programming languages through a single CLR, all .NET supported languages offer the same full platform access and full support for object-oriented programming, including inheritance and polymorphism. It is no longer necessary for host applications to provide scaled down object models specifically for scripting. The .NET framework implements classes through the use of "Namespaces" and "Assemblies." In .NET terms, an assembly is any collection of classes and resources, whether private or shared, that has been compiled into MSIL. Thus, any .NET application compiled into a MSIL library or executable is also known as an assembly. .NET assemblies are conceptually similar to JAVA packages.

The intrinsic foundation classes of the.NET framework are bundled into approximately 30 separate assemblies (DLL files in this case) loaded onto the host system during installation of the .NET framework. Each of these assemblies is then available to be referenced in code through the inclusion of its namespace. (The two top-level intrinsic namespaces of the .NET framework are "Microsoft" and "System." Most applications routinely reference members of the System namespace. The Microsoft namespace, dealing with more esoteric events and functions, is referenced less often.) Thus, for example, to implement the Show method of the MessageBox class included in the System.Windows.Forms.dll assembly, a Visual Basic.NET application might begin by importing the System.Windows.Forms namespace as shown in Listing 6.

Listing 6: A simple message box application in Visual Basic.NET.

```
Imports System.Windows.Forms

REM Module MyApp
REM Houses the application's entry point.
Public Module MyApp

   REM Main is application entry point.
  Sub Main( )
     System.Windows.Forms.MessageBox.Show_
       ("Press OK to Finish.")
  End Sub

End Module
```

Note that classes of a given assembly (including applications compiled to MSIL) can be explored using the command line MSIL disassembler utility, ildasm.exe, provided with the .NET framework software development kit (SDK)—thus, "ildasm.exe System.Windows.Forms.dll."

While the .NET framework does preserve the concepts of scripting engines and scripting hosts, the details are changed quite a bit. Most notably, .NET compiles all code, even script, through two stages before running. The first compilation stage is from application-native language (e.g., Visual Basic, JScript, and C++) to MSIL. The MSIL version of an application is CPU-portable and must be further compiled by a JIT process into native machine code before execution. The second, JIT compile to native machine code is handled within the .NET framework, is transparent to the programmer and can largely be ignored in the context of this discussion. The first compilation to MSIL has more significant implications for script authors.

For script applications (e.g., ASP.NET), compilation to MSIL is made by the new .NET scripting engines. This means that when a host relies on a scripting engine, no explicit compile step is required. It also means that compile errors (e.g., basic syntax errors) are reported separately from runtime errors and before application execution has begun, a definite benefit when debugging complex applications. It also means that performance can be detectably different the first time a script application is run after code changes. The scripting engine and host application determine automatically whether the code has been altered since last compile. If so, a compile to MSIL is made before execution. If not, the previously compiled version is used (saving time). Also, all languages compile to the same version of MSIL, requiring more commonality among the languages regarding data typing, passing of parameters, and runtime (intrinsic) functions. These changes, which have been implemented in part to reduce chance of memory leaks, overflows, and other hazards to execution safety and security, have resulted in a number of Visual Basic language changes, e.g., strong data typing (no more variants), explicit declaration of all variables, procedure parameters passed by value by default rather than by reference as previously, changes to or replacement of familiar functions, and modification of object models.

Finally, while scripting engines for both JScript and Visual Basic ship with the .NET framework, at this time only one scripting host, that for IIS (i.e., for ASP.Net) is included. There is as yet no .NET equivalent to the WSH. In the meantime, however, the .NET framework does include command-line compilers for both Visual Basic and JScript, and this is sufficient to allow hard-core script authors to still use their favorite plain text editor to program Windows using Visual Basic or JScript (as opposed to authoring in Visual Studio.NET or Visual Studio for Applications). Before running such scripts, however, they must be compiled to MSIL. For Visual Basic code this is accomplished by using vbc.exe, e.g., for the script in Listing 6,

```
vbc.exe/t:winexe/r:System.Windows.Forms.dll
  MyScript.vb
```

where "MyScript.vb" is the name of the file containing the code in Listing 6. Note the /r command line switch, which is used to explicitly point to the location of an imported namespace.

How ASP.NET Generates Client-Side Script

With the .NET framework, Microsoft also has significantly changed the preferred way for supporting client-side scripting of Web browsers. As an alternative to directly writing script for conventional HTML pages, ASP.NET has been enhanced to automatically generate client-side script as required for most common client-side scripting tasks. In particular, ASP.NET takes care of customizing client-side scripts for kind and version of Web browser being used to access ASP pages. This promises to save Web developers time and effort, though of course there will still be times when programmers will want to write their own explicit client-side scripts.

ASP.NET accomplishes automatic client-side script generation as needed through its implementation of ASP.NET "Server Controls." These controls reside in the System.Web.UI.HTMLControls and System.Web.UI.WebControls namespaces. Items in the HTML controls collection correspond to HTML elements but are augmented with additional features, and functionality (e.g., form <input> elements can be configured to automatically retain user modified values during self-referencing calls to the ASP.NET page). Items in the Web controls collection include user interface classes analogous to traditional Visual Basic classes (e.g., the text box class, which is at once analogous to the Visual Basic text box class and to HTML <input> elements of type text). The Web controls collection also contains classes of advanced functionality, e.g., data bound controls (DataGrid, DataList, Repeater), a calendar control, an ad rotator control, and several form validation controls (CompareValidator, RangeValidator, RegularExpressionValidator, RequiredFieldValidator, etc.).

Validation controls in particular tend to generate client-side script automatically. For instance, a script author of an ASP.NET page (a typical file extension is .aspx) can associate a RequiredFieldValidator control with an HTML input control of type text. This will result in the generation of client-side script (appropriate for the Web browser), which will prevent submittal of the HTML form until the user has entered data in the HTML input box. The client-side script is generated entirely by the ASP.NET host based on attributes of the relevant controls, before the generated HTML page is sent to the browser.

Web Services

The advent of the .NET framework has also provided an opportunity to reimplement from the ground up scripting practices and procedures that have grown up somewhat helter skelter since the introduction of VBScript in 1996. As an example in the area of Web applications, Microsoft has revamped earlier work on remote scripting to better follow community standards such as XML and SOAP (Clinick, 2001). The end result is an emphasis in the .NET framework on tools and facilities that facilitate the creation and implementation of "Web services." Web services, a community-driven technical approach to Web application design, continues the steady move of Web applications from the static to the dynamic. When first introduced, the Web was mostly about serving static data formatted in HTML. Over the years, Web functionality has become increasingly dynamic through extensions of both Web server and Web browser functionality. Initiatives such as Microsoft's Remote Scripting showed how Web applications could even be distributed between client and server (e.g., by allowing client-side script to instantiate and use objects running on the server). .NET Web services as implemented by Microsoft do much the same thing as remote scripting, but by making better use of new available community standards such as XML and SOAP. Web services in ASP.NET are implemented through special kinds of ASP pages (the default file extension is .asmx). These pages include a WebService directive, which informs the ASP.NET service regarding service implementation programming language and class names. Available public methods and resources are then enumerated and scripted in the .asmx page .NET Web services have advantages over older style remote scripting approaches in that they rely less on proprietary technologies and they utilize HTTP POST (rather than HTTP GET, which is limited in terms of data size communicated). By designating a particular file extension for Web services pages, ASP.NET also facilitates development by separating conventional ASP pages from those intended to be used for Web services. If you request the URL of an .asmx page from your Web browser, ASP.NET will generate and send you an HTML page suitable for testing and debugging the functionality of your Web service.

CONCLUSION

VBScript has been and continues to be an important programming language. While simple and easy to implement and having relatively low overhead, it is a rich, powerful, and modern structured programming language with many of the best features of object-oriented programming languages included. As a kind of VBA for the Web it has been particularly important as the scripting language of choice for the development of server-side ASP pages run in the context of Microsoft's IIS Web server. Though less ubiquitous and general purpose (because its not browser neutral) than JavaScript it can be useful in certain circumstances for client-side scripting of the Internet Explorer Web browser. Since 1998 and the introduction of the WSH, it has been important as a powerful, robust, shell-like language for the Microsoft Windows operating system. Microsoft and others continue to release additional scripting libraries and object models—e.g., Microsoft's recently released Windows management information scripting library that facilitates creation of operating system information gathering and administrative scripts (Stemp et al., 2002).

On the other hand, the introduction of the more unified programming environment of the .NET framework clearly suggests that VBScript is gradually being supplanted by the reunified Visual Basic.NET version of the language. With the concurrent release of ASP.NET, VBScript is no longer the preferred language for creation of new ASP applications. Even for scripting of the operating system, the more comprehensive array of classes intrinsic to the .NET framework combined with a command line utility that can compile to MSIL, makes Visual Basic.NET the better choice in many instances. VBScript

will continue to be important for some time to come, but we can anticipate that it will eventually be replaced completely by the new unified Visual Basic.NET version of the language.

GLOSSARY

Dynamic-link library (DLL) A library module containing functions and data that can be used by other modules (application or DLL).

Microsoft Scripting Runtime Implemented as a DLL, the scripting runtime makes available to script authors essential classes that are used to access and manipulate file system directories and text files, to create and retrieve key delineated lists of values, and to process encoded script modules.

.NET framework A set of extensible object classes providing comprehensive access to events and services of the Microsoft Windows platform. It rationalizes and integrates support for a wide range of languages including the new Visual Basic.NET.

Object class Encapsulates a module of reusable code and associated data structures and defines the interfaces (e.g., events, methods, and properties as used in VBScript) necessary to integrate objects of the class into an application. Object classes are used as templates from which object instances are instantiated for use in a specific program or script.

Object event An occurrence during script execution that triggers the invocation of an object event handler, i.e., a function or procedure. Events may be explicitly raised by a line of script or may fire in response to user action or an interaction with the scripting host or another object.

Object method A discrete procedure or function associated with an object. A script invokes or calls an object method at the point in code where it needs the object to perform a particular task.

Object model A road map describing the objects and associated events, methods, and properties that are part of a particular programming library or scripting host application.

Object-oriented programming language A language that supports the use of objects, including encapsulation, polymorphism, and inheritance. VBScript supports the use of objects and encapsulation, but it does not implement polymorphism and inheritance. Visual Basic.NET does.

Object property A value or attribute associated with an object instance that can be read and/or set. Most object properties are mutable during the life of the object and may be altered after the object is instantiated. Other object properties are fixed when the object is instantiated.

Scripting engine A Windows component (program) that runs scripts and provides the interfaces defined by Microsoft for scripting engines. A scripting host instantiates an instance of a scripting engine to run a script invoked in the context of the host application.

Scripting host Any application or program that provides context and environment for running scripts and provides the scripting host interfaces defined by Microsoft. When processing a script in the context of a host application, the scripting host determines the appropriate scripting engine for language used, invokes it, and calls on that engine to run the script requested.

Structured programming language A language that supports code blocks (i.e., ordered sequences of code statements that are executed and referenced as a whole), hierarchical nesting of code blocks, user-defined subroutines and functions, and the distinction between local and global variables. VBScript is a structured programming language.

Variant The only data type used in VBScript for script variables and named constants. Variants may contain many different kinds of data.

CROSS REFERENCES

See *Visual Basic; Visual C++ (Microsoft)*.

REFERENCES

Childs, M., Lomax, P., & Petrusha, R. (2000). *VBScript in a nutshell*. Sebastopol, CA: O'Reilly.

Clinick, A. (1999). Remote scripting. Retrieved June 21, 2002, from http://msdn.microsoft.com/library/en-us/dnclinic/html/scripting041299.asp

Clinick, A. (2000). Windows Script Host 5.6. Retrieved June 21, 2002, from http://msdn.microsoft.com/library/en-us/dnclinic/html/scripting11132000.asp

Clinick, A. (2001). Remote scripting in a .NET world. Retrieved June 21, 2002, from http://msdn.microsoft.com/library/en-us/dnclinic/html/scripting11122001.asp

Kiely, D. (2001). Birth (and growth) of a language. *Visual Basic Programmer's Journal*, *11*(9), 44–55.

Microsoft Corporation (2002a). Microsoft Developers Network Library. Retrieved June 21, 2002, from http://msdn.microsoft.com/library/

Microsoft Corporation (2002b). Visual Basic Scripting Edition: Version information. Retrieved June 21, 2002, from http://msdn.microsoft.com/library/en-us/script56/html/vtoriVersionInformation.asp

Microsoft Corporation (2002c). Visual Basic Scripting Edition: VBScript coding conventions. Retrieved June 21, 2002, from http:// msdn.microsoft.com/ library/en-us/script56/html/vbsCodingConventions.asp

Rahmel, D. (1999). *Visual Basic 6: Programmer's reference* (2nd ed.). Berkeley: Osborne/McGraw–Hill.

Rahmel, D. (2002). *.NET Framework: Programmer's reference*. Berkeley: Osborne/McGraw–Hill.

Stemp, G., Tsaltas, D., Wells, B., and Wilansky, E. (2002). *WMI Scripting: The missing manual* (Part 1). Retrieved June 21, 2002, from http://msdn.microsoft.com/library/en-us/dnclinic/html/scripting06112002.asp

Visual C++ (Microsoft)

Blayne E. Mayfield, *Oklahoma State University*

INTRODUCTION

From somewhat humble beginnings, Microsoft Visual C++ has evolved into a powerful suite of tools for the development of Windows applications. Software development in Visual C++ centers on the IDE (Integrated Development Environment), which provides the programmer with tools including source and resource editors, debuggers, browsers, profilers, and the MFC (Microsoft Foundation Class) library. As its name implies, MFC is a library of C++ classes that supports a broad spectrum of programming tasks for the Microsoft Windows operating systems. In addition to describing briefly the major components of the IDE, this chapter focuses on using these tools to develop applications for the Internet. Although this is a big enough topic to merit its own book, the chapter provides a foundation for further reading, understanding, and experimentation.

This chapter assumes that the reader has a general understanding of Internet applications, such as the client–server model; Comer and Stevens (1997) is a good source of information on these topics. It also assumes that the reader has a good working knowledge of object-oriented techniques (classes, objects, message passing, and inheritance) as defined and used in standard C++ applications in DOS, UNIX, or other operating systems.

The discussion and examples given below are based on Visual C++, version 6.0. As this chapter is being written a new version——Visual C++.NET——has just arrived on the development landscape. However, it is anticipated that the examples provided here should work using either version, and they should run under Windows 98, ME, NT, 2K, and XP.

THE VISUAL C++ INTEGRATED DEVELOPMENT ENVIRONMENT

Workspaces, Project Files, and the General IDE Layout

Applications in Visual C++ are maintained as a hierarchy of files. The top level of this hierarchy is the *workspace;* a workspace is a collection of one or more projects. A *project* is a collection of user-defined files, along with system-defined files that maintain information about the project. Like makefile in UNIX, a Visual C++ project is intended to ease the care and feeding of applications consisting of numerous source and header files. Also like makefile, projects keep track of file changes so that only the source and header files that have been modified are recompiled at build time. Further, as one moves from console applications into the more complex world of Windows applications, the project also help keeps track of details and changes related to the GUI (graphical user interface) aspects of the application. Generally, a workspace contains only one project, though it can contain multiple projects.

Figure 1 illustrates a typical project and various aspects of the IDE as they appear in Windows XP. Area 1 is the *toolbar* area; as with most Microsoft products, this area can be customized to meet the user's needs. Below the toolbars and on the right (area 2) is the *source editor*. Like many modern source editors, it is color-coded; i.e., keywords appear in one color, comments in another color, etc. The programmer can customize the colors and font faces used in the source editor.

Area 3 of the IDE is especially significant in its utility to the developer. This area presents different views of the

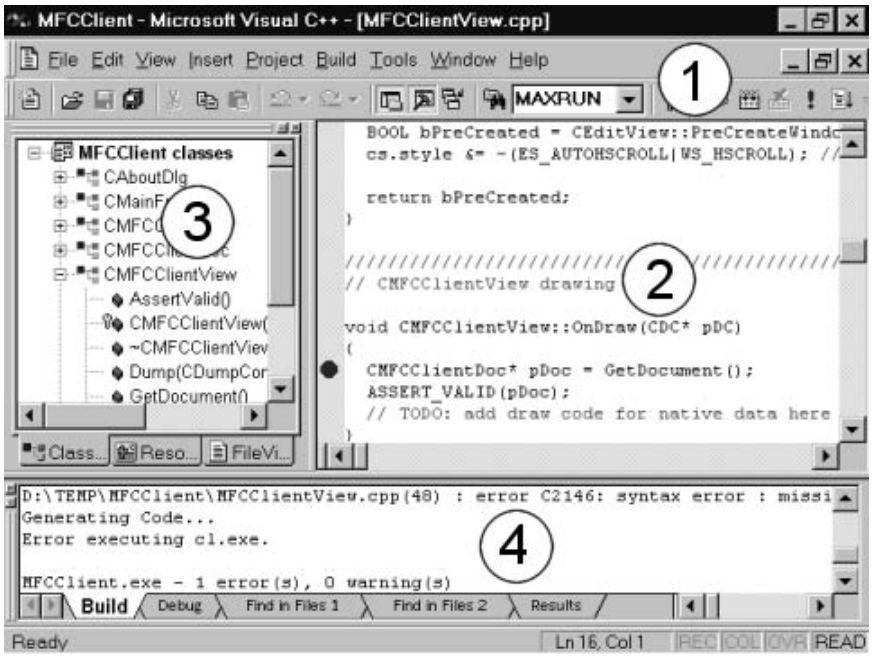


Figure 1: Commonly seen aspects of the Visual C++ IDE.

project, and one chooses the view from among the view tabs at the bottom of this area. The information in all views is presented in a tree format, similar to that used by the Windows Explorer. The three most common views are as follows:

The *file view:* The project is presented as a categorized list of all files. Double-clicking on a file name opens the file in the source editor.

The *class view:* The class and global definitions found in the application are presented. (This is the view shown in Figure 1.) Double-clicking on a class name opens in the source editor the file containing the class interface. Double-clicking on a class method name opens in the source editor the file containing the method implementation and places the cursor at the beginning of that implementation.

The *resource view:* This view presents the GUI and other Windows-specific resources associated with the application. Resources include such things as bitmaps, dialog boxes, and menus. Double-clicking on a resource opens it using the appropriate editor in the area normally occupied by the source editor. For example, Figure 2 illustrates a dialog resource open in the *dialog resource editor*. Note the toolbar at the left containing drag-and-drop icons for a host of dialog controls, such as edit boxes, buttons, combo boxes, and calendars.

The *output windows* (area 4 in Figure 1) have a multitude of uses. When one builds (i.e., compiles and links) a project, error and warning messages appear in this area. When one double-clicks on any of these messages, the file containing the offending statement is opened in the source editor, and the cursor is placed at the statement.

Debugging Tools

A Visual C++ ".exe" project commonly is built in one of two configurations: release and debug. A debug build produces a larger executable file that runs more slowly, but the executable contains information that can be used with the interactive debugging facilities of the IDE. A release build is smaller and faster, but does not support use of the interactive debugger.

The Visual C++ interactive debugger is a great asset to the development of complex programs. It permits one to set breakpoints on source code statements such that execution stops just before the marked statements are to be executed. (Figure 1 illustrates a breakpoint, which appears as a large red dot in the left margin of the marked statement.) When a breakpoint is encountered, execution of the program stops and control is turned over to the programmer. The programmer then can examine and change variable values, single step through the code, execute to a given point in the code, and continue execution without further breaks.

In a similar fashion, if a run-time error occurs, the debugger gives the programmer the option of going to that point in the program, examining and changing variables, seeing a function call trace, and continuing program execution.

One weakness of the Visual C++ debugger is the difficulty of using it to debug multithreaded programs. Regrettably, this includes many Internet applications.

Although Visual C++ is a great platform for developing many types of applications, its greatest strength is in the development of Windows GUI applications. Creating and maintaining GUI applications under Windows can be grueling. Fortunately, Visual C++ provides three tools that make this task less tedious: the *AppWizard* and *ClassWizard* and the *MFC* (discussed at length in the next section).

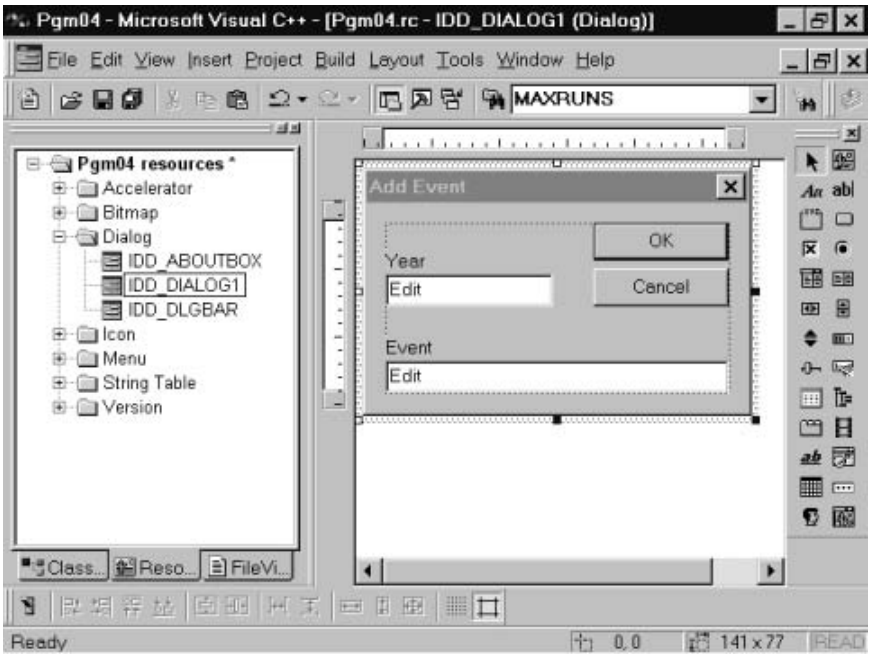


Figure 2: The dialog editor.

The AppWizard and the ClassWizard

The AppWizard is an application framework generator. That is, its function is to generate working, skeletal GUI applications. The developer is prompted through a sequence of wizard dialog boxes that present choices relating to menu items, toolbars, database support, and other commonly required application components and decisions. When the AppWizard has completed its interview of the programmer, it creates a Windows GUI application that embodies the requested features. That application can be compiled and run without the developer writing a single line of code. However, the skeletal application does not really do much; it is up to the programmer to customize the application to fulfill its functionality.

The ClassWizard is a tool that works hand in hand with the AppWizard. The function of the ClassWizard is to help the developer create new classes, especially those derived from MFC classes. The ClassWizard generates files that contain the class interface and skeletal method implementations. The ClassWizard also takes the tedium out of defining and "wiring" Windows message handling routines and managing data members related to Windows controls. For example, one can use the ClassWizard to define the effect of clicking a mouse button in a particular area of a window.

THE MICROSOFT FOUNDATION CLASS LIBRARY

MFC Overview

The MFC library of Visual C++ is a robust collection of classes that support many areas of the Windows API (application program interface). These include such things as windows, dialog controls, databases and record sets, and Internet programming. There are numerous books that provide extensive tutorial and reference coverage of MFC programming in Visual C++ (e.g. Brain & Lovette, 1999; Kruglinski et al., 1998; and White, 1999).

MFC is a strange hybrid of object-oriented and older technologies. One of its greatest weaknesses (and strengths) is its backward-compatibility to the Microsoft SDK (Software Development Kit), the non-object-oriented predecessor of MFC. To make MFC work with SDK legacy code, many non-object-oriented compromises were required in MFC. This makes MFC a sometimes confusing and frustrating patchwork of both traditional and object-oriented approaches.

Pointers to MFC objects are used commonly throughout Windows programming, so many of the methods found in MFC classes are defined as virtual functions. For example, the `CButton` class, which represents a dialog pushbutton, can be instantiated directly. But the developer might want to control how some buttons will be appear when drawn. The programmer derives a new class from `CButton` and implements in it `DrawItem`, a virtual function also defined in `CButton`. Through the miracle of virtual functions and pointers, the appropriate implementation of `DrawItem` will be called for all buttons.

Some MFC classes are abstract, requiring the developer create their own customized versions of the classes. Examples of MFC abstract classes are the document and view classes.

The Document/View Architecture

Many MFC applications are written using the *document/view architecture* as defined by Microsoft. An application that implements the document/view architecture has at its core instances of classes that are derived from four MFC classes: `CWinApp`, `CFrameWnd`, `CDocument`, and `CView`. The messaging relationship among the instances of these classes is shown in Figure 3.

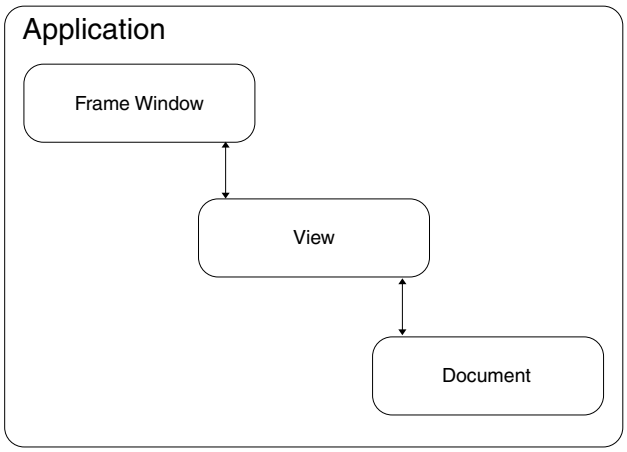


Figure 3: Document/view architecture relationships.

The *application instance* controls the application and the interaction of the other components. The *frame window* is the main window, which includes the title bar, borders, menu bar, tool bars, status bars, and the client area. Thus, the frame window is the visual interface to the user. The *view instance* generally equates to the client area of the frame; the *client area* is the "white" area of a window below the menu and tool bars. The *document instance* is responsible for managing the data files and other data stores of the application. One of the main responsibilities of the view is to act as a communication conduit between the frame window and the document; i.e. the view takes data from the document and displays it in the frame window, and it takes user input from frame components such as edit boxes and menu items and uses the input to instruct the document on data manipulation.

Single-Document and Multiple-Document Interfaces

Applications that follow the document/view architecture can be categorized by the number of documents (and, therefore, document instances) they can manipulate simultaneously. SDI (single-document interface) applications can have only one document open at a time. An example SDI application is Microsoft Paint; it can have only one image open at a time. Microsoft Word is an example of an MDI (multiple-document interface) application; conceptually, MS Word can have any number of documents open at one time. The developer chooses whether an application is SDI or MDI when running the AppWizard. In both SDI and MDI applications, a document instance can have one or more view instances associated with it so that the document data can be viewed in different ways.

WINSOCK PROGRAMMING

A Little History

The WinSock API (Windows Sockets Application Programming Interface) specification has at its core structures and functionality founded on the sockets network programming model developed at the University of California at Berkeley, plus extensions that make sockets work with the Microsoft Windows messaging system. The first widely accepted version of the specifications was WinSock 1.1 (Hall, Towfiq, Arnold, Treadwell, & Sanders, 1993), which was released in 1993 to work with Microsoft Windows 3.*x* and later versions of Windows. Windows 3.*x* was a 16-bit operating system with no support for multithreading, so certain limitations had to be observed in WinSock 1.1. In 1996 the specifications for WinSock 2 were released, and they have since been refined into the current WinSock 2.2.2 specifications (WinSock Group, 1997). Beginning with WinSock 2, support for the 16-bit versions of Windows was discontinued. The specifications were developed by the WinSock Group, made up of people from many companies and educational institutions.

WinSock supports several types of sockets including *stream sockets* (for use with TCP), *datagram sockets* (for use with UDP), and *raw sockets* (for accessing low-level communication protocols). There are two DLLs (dynamically linked libraries) that contain the library code defining WinSock: `WINSOCK.DLL` (16-bit) and `WSOCK32.DLL` (32-bit). Discussion here focuses on stream sockets using the 32-bit WinSock 2.

WinSock Without MFC

The constructs and functionality found in `WSOCK32.DLL` are not based on MFC, but rather on the Windows SDK API (Quinn & Shute, 1995). The main data type provided by this API is `SOCKET`, which is used to represent a socket handle. Briefly, in a typical client–server application that uses the WinSock 2 SDK API, the server program will contain this basic sequence of events:

1. Initialize the WinSock DLL by calling the function `WSAStartup`;
2. Declare a `SOCKET` variable;
3. Call the function `socket`, which creates a socket and returns its handle;
4. Assign the handle to the `SOCKET` variable created earlier;
5. Bind the socket to an IP address and port number by calling the `bind` function;
6. Enable the socket to take connections by calling the `listen` function;
7. Wait for an incoming connection to the socket by calling the `accept` function;
8. Send and receive data to and from the client program using the send and `recv` functions, respectively;
9. Close the socket by calling the `closesocket` function when data exchange has been completed; and
10. Terminate the WinSock `DLL` by calling the function `WSACleanup`.

It is important to note that the sequence given above describes synchronous operation. The `accept` function used in step 7 is a blocking call, meaning that further program execution is blocked while the program waits for a connection; this can lead to undesirable side effects. As will be discussed and demonstrated later, the solution to the blocking problem is to make the program

multithreaded and to place the `accept` call in a secondary thread (i.e., some thread other than the main thread).

The client program that goes along with the server program described above shares most of its same steps; specifically, to transform the sequence into a client program one simply replaces steps 5–7 with a single new step: call the function `connect` for the socket, indicating the IP address and port number to which the connection is to be made.

WinSock-Related MFC Classes

To make WinSock 2 integrate better with MFC, Microsoft has provided two MFC WinSock classes: `CSocket` and its base class `ASyncSocket`. The latter class mostly is an MFC wrapper around a `SOCKET`, but it also provides callback capabilities. `CSocket` is at a higher level of abstraction than `ASyncSocket`, and it is a blocking class. A `CSocket` object generally is used along with objects of two other MFC Classes——`CSocketFile` and `CArchive`—— to simplify socket-based communication. The use of these classes will be demonstrated in an example client–server application later.

Multithreaded Programming Issues

As mentioned earlier, a synchronous WinSock 2 program contains a function call that blocks further execution in the program until a connection has been made. In MFC GUI programs, this is particularly problematic; while execution is blocked, the user cannot communicate with the user interface. For example, if a WinSock program is blocked, the user can click on the "x" in the upper-right corner of the main program window to close it, but the program cannot and does not respond to the user interaction. For this reason, MFC WinSock programs almost always are multithreaded. By placing the blocking call in a separate thread from the GUI, the GUI can continue to interact with the user (Beveridge & Wiener, 1997).

Another pleasant benefit provided through multithreaded programming is that server applications can be made to serve multiple client applications simultaneously. This is demonstrated in the example applications that are discussed in later sections.

Unfortunately, the use of multithreading and WinSock together is itself problematic. Because of the way that the Windows infrastructure is set up, complex MFC objects, such as views, documents, and sockets, cannot be shared among threads. This usually means that these objects must be translated into handles, the thread procedure must be started, and then in the thread the handles must be transformed back into their respective objects—a messy, but usable solution!

An Example WinSock Server

The best way to learn to use WinSock with MFC (and, for that matter, the best way to learn any programming technique) is by doing. In this section and the following section, an example client–server application is presented, step by step, to illustrate how one can use the `CSocket` MFC class. In this case, the client contacts the server, which then sends to the client, once every second, a string that indicates the current date and time on the server computer.

Create the Server Project

Click on "New" from the IDE menu bar, and then select the "Project" tab in the dialog box. Choose "MFC AppWizard (exe)" as the type of project, and choose a project directory for the project name, as shown in Figure 4. (For the duration of this description, it will be assumed that the project is named `MFCServer`.) Click on the "OK" button to begin the AppWizard.

In the AppWizard, choose the single-document interface (in AppWizard step 1), uncheck the "Printing and print preview" checkbox (in step 4), check the "Windows sockets" checkbox (in step 4), and choose `CEditView`

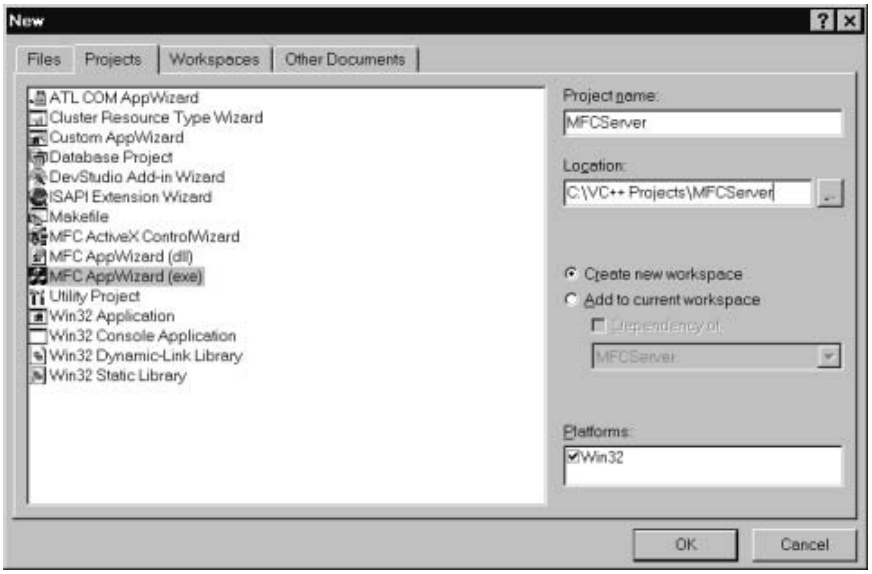


Figure 4: Creating a project.

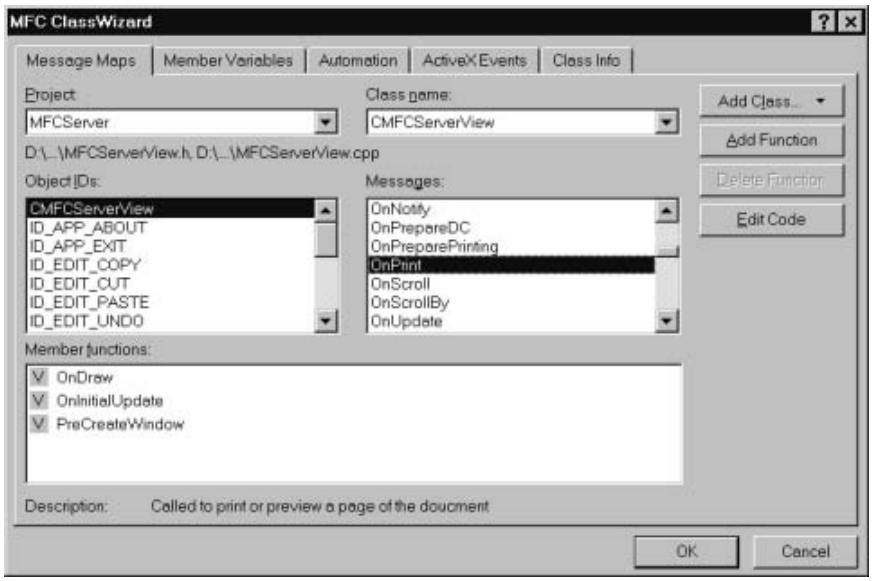


Figure 5: The ClassWizard.

from the dropdown list of view base class choices (in step 6). Accept the defaults for all other AppWizard choices.

Create a Message Handler

Invoke the ClassWizard by clicking on "View" from the IDE menu bar, and then on "ClassWizard." From the "Class name" dropdown list choose `CMFCServerView`, from the "Object IDs" list choose `CMFCServerView`, and from the "Messages" list choose `OnInitialUpdate`. The ClassWizard window should now look something like that shown in Figure 5.

Click on the "Add Function" button, and `OnInitialUpdate` appears in the "Member functions" list at the bottom of the ClassWizard window. The ClassWizard has added a declaration for this method to the view class interface, and it also has created a skeletal method implementation. Double-click the name `OnInitialUpdate` in the "Member functions" list, and the ClassWizard disappears and the cursor is placed at the newly created method implementation in the source editor. Insert the statements shown in Listing 1 after the other statements that are already in `OnInitialUpdate`. (Do not type the line numbers that are shown in the left margin of each statement; these are for discussion only.)

Listing 1: OnInitalUpdate body for example WinSock server program.

```
 (1) // Put initial message to view.
 (2) SetWindowText("Number of clients: 0");
 (3)
 (4) CSocket serverSock;
 (5)
 (6) // Create the listening socket.
 (7) if (!serverSock.Create(5555)) {
 (8)   MessageBox("Create Error");
 (9)   return;
(10) }
(11)
(12) // Activate the listening socket.
(13) if (!serverSock.Listen()) {
(14)   MessageBox("Listen Error");
(15)   return;
(16) }
(17)
(18) // This is needed to make
       multithreading work.
(19) hSocket = serverSock.Detach();
(20)
(21) // Begin the server thread,
       passing a handle to the view.
(22) AfxBeginThread(ServerThreadProc,
       GetSafeHwnd());
```

`OnInitialUpdate` is a message handler that is called automatically just before the main frame window of an application is displayed for the first time. The server displays in the application view the number of clients that it is serving currently; line 2 in Listing 1 displays zero as the initial number of clients.

Lines 4–16 set up a `CSocket` object that listens for connections from clients. The argument of the `Create` method is a port number. (A value different from the one shown can be used.) Because no host IP address is specified in the call, a value of NULL is assumed; this instructs the system to choose and assign to the socket the IP address of any available network interface on the server machine. Line 19 detaches the socket from the `CSocket` object in preparation of beginning a new thread and then stores the socket handle in the global variable hSocket for use by the worker threads. (As mentioned earlier, this is required because complex MFC objects cannot be shared between threads.) Finally, in line 22, a new worker thread is begun so that the main thread can continue to process Windows messages. The new thread serves the date/time strings to the clients. The first argument of `AfxBeginThread` is the name of the worker thread function; the second argument is a value to be passed as an

argument to the thread function. Because the thread function needs to update the current number of clients displayed in the view, it needs access to the MFC `CEditView` object. But—just as in the case of a `CSocket` object—the view object cannot be shared between threads; this is the reason for the call to `GetSafeHwnd`, which returns a safe handle for the view object.

Create the Worker Thread Procedure

Insert the implementation of the global function `ServerThreadProc` shown as Listing 2 into the file `MFCServerView.cpp` just above the implementation of `OnInitialUpdate`. The purpose of this function is to wait for a client connection, and once the connection has been made, to serve date/time strings to the client once each second. Further, after the connection has been made the function starts another thread (using the same thread function) so that other clients can be served simultaneously.

Listing 2: Thread procedure from example WinSock server program.

```
(1) UINT ServerThreadProc(LPVOID pParam) {
(2)   CSocket clientSock, serverSock;
(3)
(4)   // Retrieve view ptr so messages can
        be displayed.
(5)   CWnd *pView = CWnd::FromHandle((HWND)
        pParam);
(6)
(7)   // Attach the global socket handle
        to the local
(8)   // listening socket object.
(9)   serverSock.Attach(hSocket);
(10)
(11)  // Wait for a client connection.
(12)  if (!serverSock.Accept(clientSock)) {
(13)    AfxGetMainWnd()-> MessageBox
          ("Accept Error");
(14)    // Close the socket.
(15)    serverSock.Close();
(16)    // Return a non-zero code.
(17)    return 1;
(18)  }
(19)
(20)  // Display the current num of
        clients to the view.
(21)  CString msg;
(22)  msg.Format("Number of clients: %d",
        ++NumOfClients);
(23)  pView->SetWindowText(msg);
(24)
(25)  // Now the socket handle should be
        detached from
(26)  // from the listing socket object.
(27)  serverSock.Detach();
(28)
(29)  // After the Accept call unblocks,
        begin another
(30)  // thread to handle multiple
        clients.
(31)  AfxBeginThread(ServerThreadProc,
        pParam);
(32)
(33)  // Create a socket file and a
        storing archive.
(34)  CSocketFile sockFile(&clientSock);
(35)  CArchive archiveSend(&sockFile,
        CArchive::store);
(36)  // Send the time once a second until
        the client
(37)  // connection terminates.
(38)  TRY {
(39)    time_t tTime;
(40)    char buffer[64];
(41)    while (1) {
(42)      // Get current time from the
            server machine.
(43)      time(&tTime);
(44)      // Format and send the server
            time to the client.
(45)      strcpy(buffer, ctime(&tTime));
(46)      buffer[strlen(buffer)-1] = '\0';
(47)      sockFile.Write(buffer, strlen
            (buffer)+1);
(48)      // Now wait a second...
(59)      Sleep(1000);
(60)    }
(51)  }
(52)  CATCH (CFileException, pEx) {
(53)    // Communication was lost with
          the client;
(54)    // update the view.
(55)    msg.Format("Number of clients: %d",
(56)      --NumOfClients);
(57)    pView->SetWindowText(msg);
(58)    // close the client socket.
(59)    clientSock.Close();
(60)  }
(61)  END_CATCH
(62)
(63)  // Return success code.
(64)  return 0;
(65)  }
```

The first couple of actions taken in `ServerThreadProc` are there to recover the MFC objects from the handles created in the main thread. Line 5 in Listing 2 retrieves the view from the handle that was passed as an argument to the thread function. Line 9 attaches the listening socket (the handle of which is stored in the global variable `hSocket`) to a local `CSocket` object.

In line 12 the listening socket waits for a connection from a client; as mentioned earlier, a call to `Accept` is a blocking call, so execution is suspended until a connection is requested by a client. Once the call unblocks, lines 21–23 update the view to reflect that another client is being served. Line 27 detaches the listening socket from the local `CSocket` object in preparation for line 31, which starts another thread to accept another client connection.

Lines 34–61 of Listing 2 do the actual serving of strings to the client. First, lines 34–35 set up a

CSocketFile object using a CArchive to simplify communication with the client. Then a loop is entered in which the date/time is retrieved, formatted as a string, and sent to the client. At the end of the loop, the thread is put to sleep for 1 s (1,000 ms) before the next string is sent. This loop will continue until the user terminates the server program or until the connection to the client is closed. Note the use of TRY, CATCH, and END_CATCH; these are macros that implement MFC's own form of exception handling, which is used by CSocket and its parent class CAsyncSocket. The use of TRY and CATCH blocks in the thread function keep the program from terminating abnormally when the connection to the client is broken.

Final Changes

The last modification to the project is to add a #include and to declare global variables that will be used by OnInitialUpdate and ServerThreadProc. Insert these statements following the default #include statements in MFCServerView.cpp:

```
#include <time.h>
SOCKET hSocket;
int NumOfClients = 0;
```

Build the Project

Build (i.e., compile and link) the project. Do not run it until the client program, described below, has been written and built.

An Example WinSock Client

The client program presented communicates with the server program described above. It uses many of the same ideas and techniques as the server. Before beginning work on the client program, make sure to close the server program workspace.

Create the Client Project

Click on "New" from the IDE menu bar, and then select the "Project" tab in the dialog box. Choose "MFC AppWizard (exe)" as the type of project, and choose a project directory for the project name. (For the duration of this description, it will be assumed that the project is named MFCClient.) Click on the "OK" button to begin the AppWizard.

In the AppWizard, choose the single-document interface (in AppWizard step 1), uncheck the "Printing and print preview" checkbox (in step 4), check the "Windows sockets" checkbox (in step 4), and choose CEditView from the dropdown list of view base class choices (in step 6). Accept the defaults for all other AppWizard choices.

Create a Message Handler

Invoke the ClassWizard by clicking on "View" from the IDE menu bar, then on "ClassWizard." From the "Class name" dropdown list choose CMFCClientView, from the "Object IDs" list choose CMFCClientView, and from the "Messages" list choose OnInitialUpdate. Click on the "Add Function" button, and OnInitialUpdate appears in the "Member functions" list at the bottom of the ClassWizard window. Double-click the name OnInitialUpdate in the "Member functions" list to be taken to the newly created method implementation in the source editor. Insert the statements shown in Listing 3 after the other statements that are already in OnInitialUpdate.

Listing 3: OnInitalUpdate body for example WinSock client program.

```
 (1) CSocket serverSock;
 (2)
 (3) // Create the listening socket.
 (4) if (!serverSock.Create()) {
 (5)   MessageBox("Create Error");
 (6)   return;
 (7) }
 (8)
 (9) // Connect to the server.
(10) if (!serverSock.Connect("localhost",
       5555)) {
(11)   MessageBox("Connect Error");
(12)   return;
(13) }
(14)
(15) // This is needed to make
       multithreading work.
(16) hSocket = serverSock.Detach();
(17)
(18) // Begin the server thread, passing
       a handle to the view.
(19) AfxBeginThread(ClientThreadProc,
       GetSafeHwnd());
```

The major difference between this implementation of OnInitialUpdate and its counterpart in the server program is the call to the method Connect in line 10. In this example it is assumed that the server and client are running on the same computer, so "localhost" has been used as the first argument for Connect; if the server and client are on different computers, the IP address or host name of the server computer should be substituted. The second argument of Connect is a port number that matches the one used in the server program. Notice that the client program also is multithreaded, even though it will connect to only one server. Multithreading is used here so that the client program can continue to process Windows messages while it communicates with the server program.

Create the Worker Thread Procedure

Insert the implementation of the global function ClientThreadProc shown as Listing 4 into the file MFCClientView.cpp just above the implementation of OnInitialUpdate. The purpose of this function is to receive strings from the server socket and to display them in the view.

Listing 4: Thread procedure from example WinSock client program.

```
(1) UINT ClientThreadProc(LPVOID pParam) {
(2)   CSocket serverSock;
(3)
```

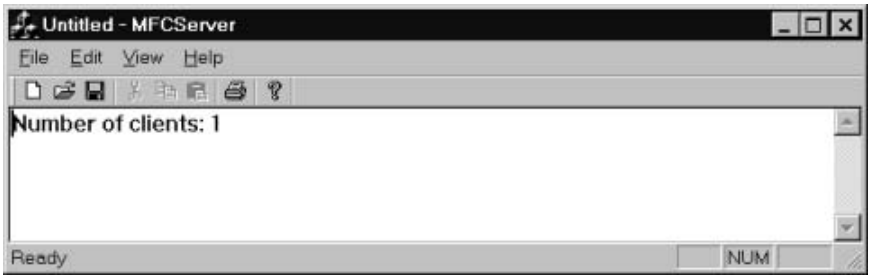


Figure 6: The MFCServer app window.

```
 (4)    // Retrieve view ptr so messages
          can be displayed.
 (5)    CWnd *pView = CWnd::FromHandle((HWND)
          pParam);
 (6)
 (7)    // Attach the global socket handle
          to the local
 (8)    // server socket object.
 (9)    serverSock.Attach(hSocket);
(10)
(11)    // Create a socket file and a
          loading archive.
(12)    CSocketFile sockFile(&serverSock);
(13)    CArchive archiveRecv(&sockFile,
          CArchive::load);
(14)
(15)    // Retrieve the time repeatedly
          until the server
(16)    // connection terminates.
(17)    TRY {
(20)      char buffer[64];
(18)      while (1) {
(19)        // Retrieve the time from the
              server.
(21)        sockFile.Read(buffer, 64);
(22)        // Display the string in the view.
(23)        pView->SetWindowText(buffer);
(24)      }
(25)    }
(26)    CATCH (CFileException, pEx) {
(27)        // Communication was lost
              with the server;
(28)        // close the server socket.
(29)        serverSock.Close();
(30)      }
(31)      END_CATCH
(32)
(33)      // Return success code.
(34)      return 0;
(35)    }
```

The important differences between this function and `ServerThreadProc` seen earlier are in lines 21 and 23 of `ClientThreadProc`. The loop body reads a string from the server connection and then displays it in the view window.

Final Changes

Only one global variable is needed to pass socket handle information between `OnInitialUpdate` and `ServerThreadProc`. Insert the following statement after the #include statements in `MFCClientView.cpp`: `SOCKET hSocket;`

Build the Project and Test the Client–Server Application

Build the client project. To test the application, first start the server program named `MFCServer.exe`, which is found in the `Debug` directory of the server project. Then start one or more copies of the client program named `MFCClient.exe`, which is found in the `Debug` directory of the client project. One should see the number of clients displayed in the server window, as shown in Figure 6, and the current time—updated once per second—displayed in each client window, as shown in Figure 7.

WININET CLIENT PROGRAMMING

An Alternative to WinSock?

WinInet (Win32 Internet) is an Internet API from Microsoft written specifically for their Windows operating

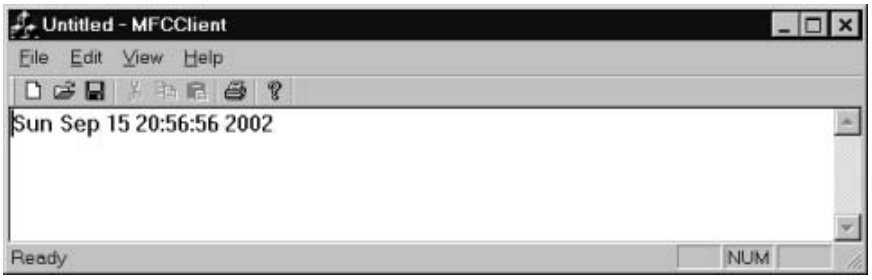


Figure 7: The MFCClient app window.

system. Kruglinski, Shepherd, and Wingo (1998) list several specific advantages of WinInet over WinSock:

File caching,
Support for proxy servers,
An easier API,
Security via authentication,
Buffered I/O capabilities, and
Better support functions.

Does this mean that WinInet is a better alternative to WinSock? No. WinInet is designed specifically for writing FTP, Gopher, and HTTP client programs; there is no support for writing server software nor for writing Internet applications that fall outside the three categories listed above. The general sequence of events for writing a WinInet client application is as follows:

Create an instance of the MFC class `CInternetSession`;
Open a connection from the `CInternetSession` instance to a URL (Universal Resource Locator) using the method `OpenURL`; if the argument to `OpenURL` is of the form "ftp://...," "gopher://...," or "http://..." the method will return a `CInternetFile` pointer, a `CGopherFile` pointer, or a `CHttpFile` pointer, respectively.
Use the appropriate file pointer to call the methods `Read`, `Write`, `ReadString`, and `WriteString` to communicate with the server.

A WinInet Example

Listing 5 presents the WinInet code segment from an MFC console application that reads and displays the HTML code that makes up the Microsoft main Web page. The file containing the code shown in Listing 5 also must contain the preprocessor statement `#include <afxinet.h>`.

Listing 5: WinInet console application code segment.

```
(1) CInternetSession session;
(2) CHttpFile *pFile;
(3) CString buffer;
(4) DWORD dwStatus;
(5)
(6) // Open the URL for Microsoft's main
      page.
(7) pFile =
(8)    (CHttpFile*)session.OpenURL(
         "http://microsoft.com");
(9)
(10) // Make sure the open was successful.
(11) pFile->QueryInfoStatusCode(dwStatus);
(12) if ((dwStatus < 200) || (dwStatus >
       299)) {
(13)    // The URL could not be opened.
(14)    return false;
(15) }
(16)
(17) // Get and display the HTML code from
       the Web page.
(18) while (pFile->ReadString(buffer))
(19)    cout << (const char*)buffer << endl;
(20)
(21) // Close the HTTP file.
(22) pFile->Close();
(23) // Close the connection.
(24) session.Close();
```

OLE, ACTIVEX, AND COM

OLE

OLE (Object Linking and Embedding) was one of Microsoft's first attempts to create a common, cross-application API format. OLE makes it possible to share data, objects, and components among different applications. One of the first and most obvious applications of OLE was the ability to drag and drop data and objects from one application (for example, Microsoft Excel) into another application (for example, Microsoft Word).

ActiveX

ActiveX is the successor to OLE. (More specifically, ActiveX controls are the successor to OLE controls.) Kruglinski et al. (1998) say that one can view ActiveX "as something that was created when the 'old' OLE collided with the Internet." Visual C++ comes ready with numerous ActiveX controls available. For example, a month calendar control is available on the toolbar of the IDE dialog editor; Figure 8 depicts the use of this control in a simple application. For those who want to write their own custom ActiveX controls, Visual C++ includes the ATL (Active Template Library).

COM

COM (Component Object Model) is the underlying architecture used in ActiveX technology. The main advantage of COM (and other component technologies) is that of component reuse. Rogerson (1997) points out that COM components can be used across many procedural and object-oriented languages, and they can be updated or relocated (locally or on a network) with little or no effect on the applications that use the components.

ActiveX and Web Programming

Numerous ActiveX controls are available to promote the development of Visual C++ Web-aware applications. A good example is the Web browser control provided by Microsoft; using it, one can write Web browser applications or applications that have Web browsers embedded within them. Because this control is based on the technology used in Microsoft Internet Explorer, one can use it to make applications have the look and feel of Internet Explorer. From the other side of the browser wars comes another Web browser control provided by the Mozilla Group (2000). This control is based on the Gecko layout engine, which is the same technology used in the Mozilla and Netscape Web browsers.

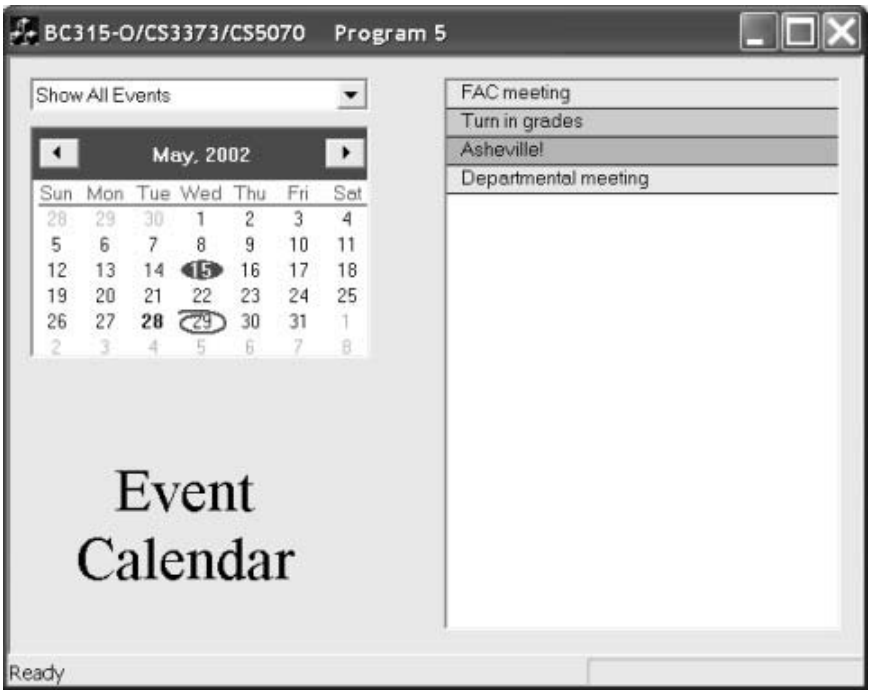


Figure 8: The ActiveX month calendar control in use.

VISUAL C++.NET: A COMPARISON

As this chapter is being written, Visual C++.NET has just recently been released. Compared to its predecessor, Visual C++ 6.0, Microsoft Corporation (2002) claims many enhancements, including the following:

The IDE has been enhanced with new features such as user-definable bookmarks within source files and the ability to highlight and then comment (or uncomment) a sequence of statements with a single mouse click.

The Visual Debugger has been enhanced. For example, one can debug multiple processes simultaneously; this is helpful in debugging client–server applications.

The ClassWizard has been replaced by options on the context menus of the class view. This is more logical than its previous location on the "View" menu of the IDE.

Support has been added for XML, XML schemas, and SOAP (Simple Object Access Protocol). Developers also have been given the ability to create and expose ATL COM components as XML Web services.

The new ATL Server, which is a successor to ASP (Active Server Pages), but which is written in C++ for greater speed and integration into C++ applications, has been added.

CONCLUSION

Visual C++ is a powerful suite of application development tools. These applications can be as simple as highly portable console applications and as complex as Windows GUI applications. Visual C++ and MFC support the development of Internet-aware applications with tools such as WinSock, WinInet, ActiveX controls, and COM components. With its latest release, .NET, Microsoft continues to expand the Internet development capabilities of Visual C++ with new features such as XML, XML schemas, SOAP, and ATL server. It is worth the effort to learn to use these tools, which can significantly increase productivity and decrease development time.

GLOSSARY

ActiveX Microsoft's follow-up to OLE, incorporating newer Internet concepts into the sharing of data, objects, and components.

AppWizard A Visual C++ tool for generating skeletal Windows/MFC applications.

ATL (Active Template Library) A suite of tools for the development of new ActiveX controls.

ATL Server A Microsoft .NET successor to ASP (Active Server Pages).

ClassWizard A Visual C++ tool for maintaining MFC-derived classes and message handlers.

COM (common object model) The underlying architecture of ActiveX.

IDE (integrated development environment) The development environment of Visual C++, including editors, variable watch windows, and class/file/resource browsers.

MFC (Microsoft Foundation Class) Library A library of classes that support the development of Visual C++ applications.

OLE (object linking and embedding) One of Microsoft's early protocols for sharing data, objects, and components among different applications.

WinInet (Win32 Internet) A Windows API for writing FTP, Gopher, and HTTP clients.

WinSock API (Windows sockets application programming interface) A WindowsAPI based on the BSD sockets paradigm.

CROSS REFERENCES

See *ActiveX; C/C++; Visual Basic*.

REFERENCES

Beveridge, J., & Wiener, R. (1997). *Multithreading applications in Win32*. Reading, MA: Addison–Wesley.

Brain, M., & Lovette, L. (1999). *Developing professional applications for Windows 98 and NT using MFC*. Upper Saddle River, NJ: Prentice Hall PTR.

Comer, D. L., & Stevens, D. L. (1997). *Internetworking with TCP/IP vol. III client–server programming and applications—Windows Sockets version*. Upper Saddle River, NJ: Prentice Hall.

Hall, M., Towfiq, M., Arnold, G., Treadwell, D., & Sanders, H. (1993). Windows Sockets: An open interface for network programming under Microsoft Windows, Version 1.1. Retrieved September 15, 2002, from http://burks.brighton.ac.uk/burks/pcinfo/progdocs/winsock/winsock.htm

Kruglinski, D. J., Shepherd, G., & Wingo, S. (1998). *Programming Microsoft Visual C++* (5th ed.). Redmond, WA: Microsoft Press.

Microsoft Corporation (2002). Visual C++. Net Home Page. Retrieved May 29, 2002, from http://msdn.microsoft.com/visualc

Mozilla Group (2000). Mozilla ActiveX Control. Retrieved May 29, 2002, from http://www.mozilla.org/docs/codestock99/mozcontrol/mozcontrol_files/v3_document.htm

Quinn, B., & Shute, D. (1995). *Windows Sockets Network programming*. Reading, MA: Addison–Wesley.

Rogerson, D. (1997). *Inside COM*. Redmond, WA: Microsoft Press.

White, D. (1999). *MFC programming with Visual C++ 6 unleashed*. Indianapolis, IN: SAMS.

WinSock Group (1997). Windows Sockets 2 Application Programming Interface: An interface for transparent network programming under Microsoft Windows, revision 2.2.2. Retrieved September 15, 2002, from ftp://ftp.microsoft.com/bussys/winsock/winsock2/WSAPI22.DOC

Voice over Internet Protocol (IP)

Roy Morris, *Capitol College*

INTRODUCTION

The voice over Internet protocol ("VOIP") is where conventional telecommunications meets the technologies used in the Internet. This chapter will begin by explaining communications theory and conventional telephony, which pre-date the popularized Internet. They serve as a foundation for the VOIP, because almost all voice communications networks must interconnect, and therefore must be backwards compatible with conventional telephony networks. The chapter will conclude with specifics of VOIP design and deployment.

COMMUNICATIONS THEORY: BITS OF TRANSMITTED INFORMATION

Basic communications theory concepts form the building block of all communications applications, and their understanding. VOIP is no exception.

"Communication" is simply the transmission of "information" from one place (or device or person) to another, over a "transmission network." "Voice communications" refers to transmitting the information contained in a voice signal. The distance traveled by a transmitted signal can be as short as from one chip to another chip on a circuit board, to across the street, to around the world, or even to the most distant galaxies in outer space.

Bits: The Most Basic Form of Information

In digital parlance, a "bit" is the smallest measure of information (Newton, 1998). It is the amount of information required to distinguish between two equally likely possibilities or choices. A bit can be used as a building block for all information or messages and/or as a measure of information. A bit is typically given one of two arbitrary logical values: e.g., 1 or 0. By stringing together bits, complex messages can be formed representing familiar message forms, such as letters, sentences, text documents, voice signals, pictures, and movies. Table 1 illustrates one arbitrary binary encoding scheme for the 26 letters of the alphabet.

Over the years, one standardized binary encoding scheme evolved for all letters of the alphabet, plus some additional characters, which is universally understood by almost all digital devices (Truxal, 1990). That text encoding scheme is known as "ASCII." ASCII has 128 standardized combinations of seven bits to represent 128 characters, including upper and lower case letters, numbers, and symbols. With such a universally understood binary text encoding scheme, we can transmit letters strung together to form large text documents that can be transmitted over digital transmission facilities, received, and ultimately displayed in distant locations. The key to the successful binary transmission of text is that the transmitter and receiver of the binary transmission must understand the same standardized vocabulary used for that text encoding—which is primarily ASCII.

Similarly, using standardized voice communication encoding schemes, voice communication signals can be represented as strings of 1s and 0s and then transmitted over data networks, received, and ultimately heard in distant locations. VOIP capitalizes on this ability, transmitting those strings of bits over a data network using the Internet protocol (IP). However, to the naked eye, all binary strings look the same—like strings of 1s and 0s. What differentiates one kind of transmitted binary string from another is a common understanding between the encoder (at the transmitter) and decoder (at the receiver) of what that string represents (e.g., text vs. voice) and how it was encoded (e.g., did it use ASCII, if it was text? Did it use a particular voice coder/decoder, if it is voice?)

Table 1 One Illustrative Encoding of Letters

A	00000	I	01000	Q	10000	Y	10000
B	00001	J	01001	R	11001	Z	11001
C	00010	K	01010	S	11010		11010
D	00011	L	01011	T	11011		11011
E	00100	M	01100	U	11100		11100
F	00101	N	01101	V	11101		11101
G	00110	O	01110	W	11110		11110
H	00111	P	01111	X	11111		11111

Transmission of Voice Signals

Almost all voice communications signals begin as analog sound signals. An analog signal is one that is continuous in shape (i.e., no sharp angles) and typically analogous in shape to the original signal or physical phenomenon it represents (Newton, 1998). The physical signals that our bodies generate (e.g., sound and visual) and receive (i.e., sense) are analog (Truxal, 1990). This is very important in the design of telephony systems, because almost all information that we are concerned with in telephony is ultimately either originated and/or received by a human being.

When we speak or play an instrument, combinations of one or more waves of alternating compression and expansion of air are generated at the sound's source, e.g., the vocal chords in our throats or the reeds of instruments. Because sound energy in air does not travel well over long distances, long-distance communications requires that these sound waves be converted to alternate forms of energy by a transducer, and then that signal energy is conveyed over a compatible transmission medium, such as an electronic transmission channel. For voice communications, the microphone of a telephone device is typically used to transform analog sound wave energy into analog electrical wave energy to be transmitted over a caller's local telephone wires.

To transform the electrical signal back to a sound wave that the human ear can detect, the analog electrical signal is applied to a "reverse microphone" transducer (such as the an earpiece of a telephone). The "reverse microphone," or speaker, outputs physical patterns of compressed air corresponding to the original sound wave. The entire transformation from analog sound wave to analog electrical signal and back to analog sound wave is illustrated in Figure 1.

The Characteristics of an Analog Voice Signal

Communications networks and their elements are designed to cost-effectively accommodate the signals they are intended to carry, with the least degradation (e.g., distortion, delay, etc). Telephony networks were optimized to carry analog voice signals over long distances using the most cost-effective technology available. However, the original designs were done from the late 1800s through the early 1900s. Those voice-centric design templates continue to influence telephony network design today due to the need for new networks to interoperate with older ones, and the need for those older networks to be backwards compatible with older equipment (which sometimes dates back to the early 1900s).

A human voice signal is made up of alternating signal components of various frequencies, mostly within the range from 100 to 3,400 alternating cycles per second. Although this range may vary somewhat from person to person, analog telephony networks were designed based on this assumption. Virtually all of the components of the telephone networks have been built to limit the frequencies transmitted to those in this predefined voiceband. Any voice frequencies outside this voiceband are filtered out by the network equipment and, in turn, not transmitted.

This voiceband range is broad enough to allow a telephone listener both to recognize the person who is speaking and to understand what he or she is saying. High fidelity devices, such as CDs and FM radio, transmit a wider range of sound frequencies (up to 10,000 cycles per second or more). With conventional telephony, any attempt to transmit sound from such devices would be filtered to only include the voiceband frequency components. However, with the flexibility of VOIP, it is at least theoretically possible to cost-effectively deploy a customized high fidelity telephony service that could coexist with voiceband VOIP services.

Digitizing an Analog Voice Signal

Most long distance transmission systems in use today (including those using VOIP) transmit voice signals digitally. To transmit an analog signal digitally, its analog version must be transformed from its electrical analog form to a digital form, thus changing its representation from a varying electrical voltage to an equivalent string of discrete 0s and 1s. This transformation is done by a digitizer, sometimes referred to as an "analog to digital converter," "A/D Converter," or "CODEC" (COder-DECoder).

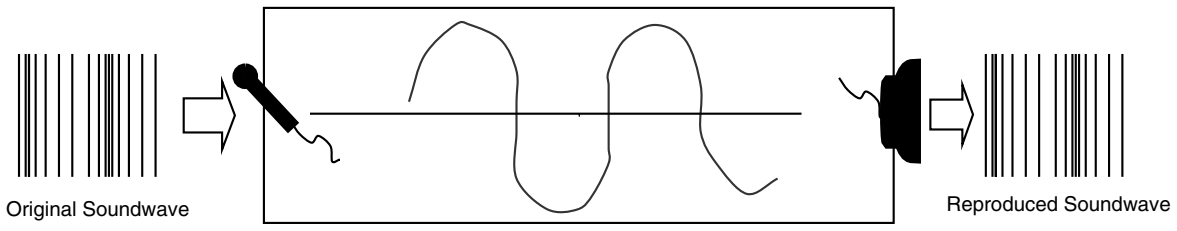


Figure 1: Transmission of a sound wave and its reconstruction.

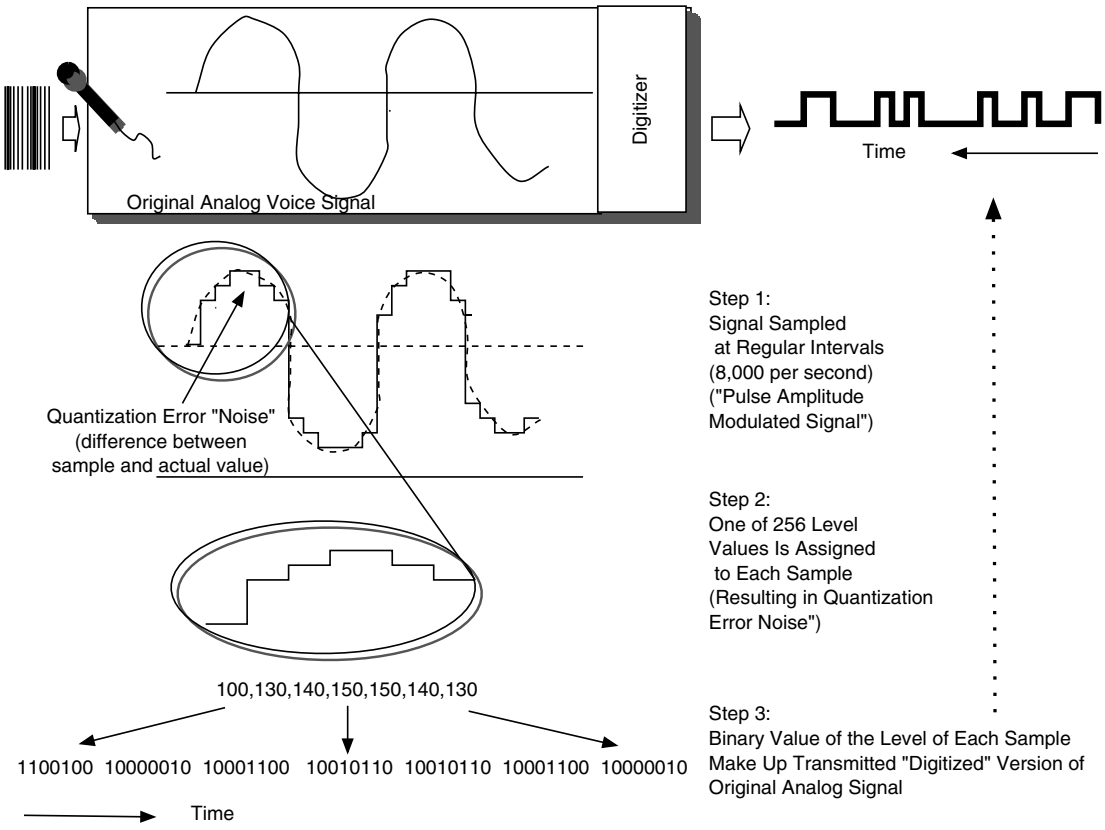


Figure 2: Digitization of a voice signal.

(Newton, 1998) and (Truxal, 1990). A digitizer takes samples of the height (i.e., amplitude) of the incoming analog signal at regular time intervals (e.g., 8,000 times per second). For each of these regularly timed measurements, the digitizer outputs a fixed length digital number (expressed in terms of logical 0s and 1s) that corresponds to the size of the signal height measured for each sample. See, e.g., Figure 2.

Depending on the digitizer, the number of bits that are used to represent each level of signal amplitude can vary. If eight bits are used to represent the measurements of each sample's height, then 256 height levels can be represented. The use of eight bits to represent each sample of a voice signal is fairly typical. The rounding error or difference between the actual height of each sample and the true height of the sampled signal at the sampled point is called "quantization error."

More bits could be used to represent each sample's amplitude, allowing more discrete levels to digitize a signal and, in turn, reducing the amount of quantization (or rounding error) noise. However, the additional bits for each sample would require additional transmission capacity with little offsetting benefit. Experience has shown that the human ear cannot easily detect the noise created by quantization error with eight bits sample encoding.

The minimum rate at which a CODEC must sample a voice signal is the Nyquist rate. Nyquist determined that in order for a sampled representation of a signal to accurately capture the entire range of frequencies in the original signal, the samples must be taken at a rate that is more than two times the highest frequency component of the original signal. Thus, in order to transmit a voice signal digitally and achieve the same performance as conventional analog telephony networks, the transmitted voiceband signal, with its highest frequency assumed to be 3,400 cycles per second, must be sampled at more than 6,800 samples per second. It is common to oversample a voice signal using a sample rate of 8,000 samples per second.

With eight bit encoding per sample, the digital equivalent of a voiceband signal that is sampled 8,000 times per second requires 64,000 bits per second (8,000 samples per second $\times$ 8 bits per sample). This is known as the G.711 standard.

CODECs often incorporate techniques for reducing the transmitted bit rate from the base amount of 64,000 bits per second. Some of those techniques include (1) using nonlinear (logarithmic) sampling levels, (2) transmitting only changes in signal levels, or (3) removing redundant information while maintaining signal quality and minimizing processing delay (Castelli, 2002). Using nonlinear sampling levels is known as "companding." Companding has the effect of decreasing the number of sample levels required by proportionately increasing the sample levels in the most relevant height range for the signal. Companding also has the secondary effect of reducing the quantization noise in the relevant amplitude range, compared to that in noncompanded digitized signals, for a given number of bits per sample. Differential pulse

Table 2 Summary of Compression Algorithm Performance (Source: Cisco, Understanding Codec Complexity (2002) and Sanford (1999))

Compression	Bandwidth (Kbit/s)	MOS score	Delay (ms)
PCM (G.711)	64	4.4	0.75
ADPCM (G.726)	32	4.2	1
LD-CELP (G.728)	16	4.2	3.5
CS-ACELP (G.729)	8	4.2	10
MPMLG (G.723.1)	6.3	3.9	30
ACLEP (G.7231.1)	5.3	3.5	30

code modulation (DPCM) is another method used to reduce the number of bits transmitted without compromising quality. DPCM only transmits the changes in levels, rather than absolute levels. Another variation on DPCM is adaptive pulse code modulation, which can halve the bit transmission requirements. Current compression algorithms can reduce the 64,000 bit per second transmission requirement for a voice signal to as low as 8,000 bits per second without incurring a substantial increase in delay.

Table 2 summarizes some of the most popular voice compression algorithms used in Internet telephony. The frame of reference is the G.711 standard. Note that its MOS ("Mean Opinion Score," an objective measurement of the perceived quality of the received voice which ranges from 0 to 5) is 4.4. Processing delay is negligible. All other compression systems are more efficient (less than 64 kb/s of bandwidth consumption per voice conversation) but exhibit additional processing delays, which may cause additional problems in long, echo-prone circuits. For wide area networks (WAN), the G.729 standard is one of the most popular, because it represents a reasonable compromise between efficiency and delay performance, at a modest 10 ms. The G.711 is most popular around the campus, where bandwidth cost is generally not an issue.

THE TELEPHONY NETWORK

The Communications Network Topology

From a user's perspective, a telephone network is a big cloud that connects the originating points for transmitted information with terminating points for information. A network is typically made up of nodes interconnecting transmission links to form transmission paths for the transmitted information to follow. The transmission links in a typical network often use different transmission modalities (e.g., wire, wireless, fiber). The nodes may contain electronic devices, such as switches, which establish a temporary or permanent path to be taken by the communicated information traveling from one link onto another link, and/or those nodes may contain conversion devices that change the modality of the transmitted signal to allow it to pass between links using different modalities (e.g., connecting a wired link to a wireless link, an analog link to a digital link, or a fiber link to a copper wire link). See Figure 3.

Transmission Links

Most communications, whether digital or analog, are transmitted over wire, fiber, or wireless links (i.e., the atmosphere). Each physical medium has its own physical characteristics that limit the ability of the transmitted

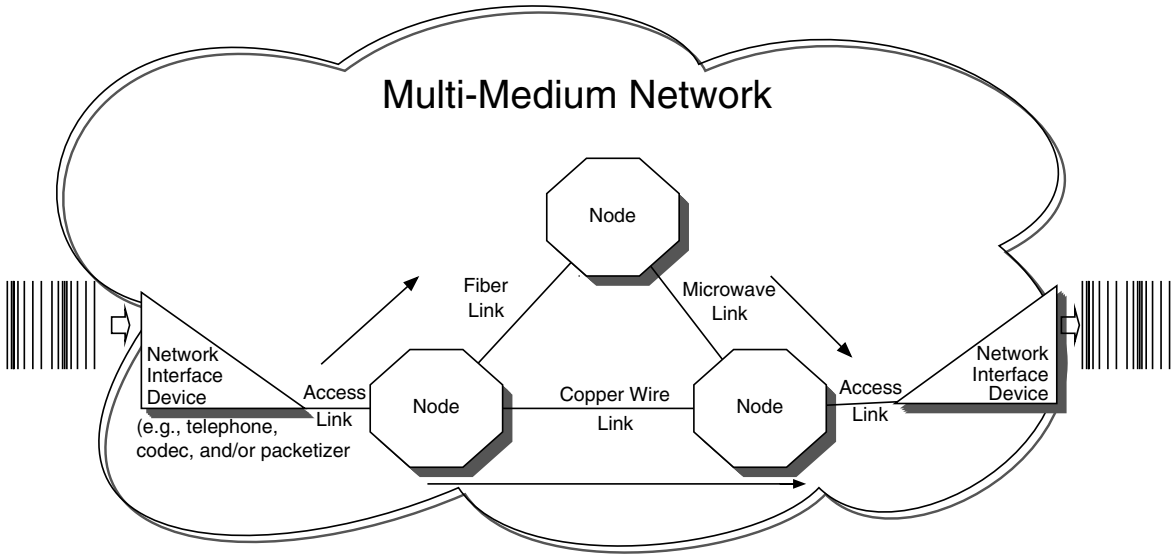


Figure 3: Multimedium network.

signal (whether analog or digitally encoded) to accurately and timely propagate.

Wire transmission includes wire pairs, coaxial cable, and even single wires (where the "return path" is the physical common ground). With wire transmission, the transducer varies the electrical pressure, or voltage, applied at the transmitting end of the transmission link in proportion to the transmitted signal, in effect pushing electrons on one wire and pulling electrons on the other wire. The electrons flow through this loop path to the node at the far end of the transmission link. At the far end of the transmission link, a transducer senses this electron movement and pressure to recreate the transmitted signal, or something corresponding to it.

Fiber links involve the transmission of communications using a light signal through a very high-quality, thin wire-like glass fiber. The signal is transmitted using a transducer that turns a light signal source (typically, a laser or light-emitting diode [LED]) at the information source end of the transmission medium on and then off or in varying intensities. Those variations are designed to be some function of the transmitted information signal. A "reverse" transducer at the far end node connected to the fiber link senses these variations in light intensity and ultimately deciphers the transmitted information.

Finally, wireless transmission links employ electromagnetic waves (such as radio and light waves) at the transmitting end that are propagated through the open atmosphere. The transmitted information is used to vary some physical characteristic of the transmitted electromagnetic signal. A receiver-type transducer at the receiving end of the open-atmosphere medium detects these variations in radio or light waves and, in turn, can recreate the original signal and the information transmitted.

Transmission System Imperfections: Noise, Loss, and Delay

Each transmission link in a network, and each node connecting those links, introduces some amount of noise, loss, and delay into the transmitted signal. These physical transmission imperfections affect both digitally and non-digitally encoded signals. However, depending on whether the signals transmitted are analog or digital, the transmission channel's physical limitations manifest themselves somewhat differently—as described later on in this chapter.

Transmission noise introduced into the transmitted signal during transmission causes the signal received at the far end of a transmission link to vary from the originally transmitted source signal. Transmission noise is typically uncorrelated with the original transmission signal. Where noise is present, detection and, therefore, reconstruction of the original transmitted signal at the receiving end is more difficult or impossible. Transmission noise, when corrupting a digitally encoded signal, can result in a receiver falsely identifying some of the 1s as 0s, and some 0s as 1s.

The ability of a receiver to accurately detect the characteristics of the original transmitted signal, and therefore to accurately reproduce it, is directly a function of the signal-to-noise ratio (S/N ratio) of the received signal. As the noise added by the transmission link becomes very large in comparison to the power of the transmitted signal, more errors in detection will occur. Boosting the power of the original signal at the transmitting end of the link can increase the signal-to-noise ratio, but this is not always possible due to concerns about such problems as cross-talk between adjacent channels, and interference with other devices and systems. Special encoding schemes can be used to allow the receiver to check and/or correct for errors caused by random noise. The particular schemes for doing this error detection and correction are not important here, except to note that they all require the "overhead" of transmitting additional redundant information (and therefore additional bits) along with the digitally encoded version of the original signal.

There is one exception to the idea that noise is a bad thing. Specifically, a limited amount of noise may sometimes be intentionally *added* to a digitally transmitted voice signal during quiet periods, when neither party is speaking; otherwise the lack of noise could lead callers to perceive that the line has been disconnected.

Transmission loss refers to the loss of signal power as the original signal traverses a transmission link. Amplifying all or selected frequencies of the received signal at the receiving end of the link cannot always compensate for signal loss. For example, in the presence of noise, if there is significant overlap in the spectral power of the noise signal and the original signal, receiver amplification will tend to amplify the unwanted noise along with the original transmitted signal. The result would be little improvement in the receiver's ability to discriminate between the now amplified original transmission signal and amplified unwanted noise and thus its ability to accurately detect the signal. The frequency range over which the loss is relatively flat is known as the bandwidth.

Transmission delay (or latency) is a measure of how much later a signal is received, as compared to when it was transmitted. Delay can create two types of problems. At the transmitted signal level, delay can vary with the spectral range of the signal. For transmitted digital pulses, this delay distortion can result in very distorted-looking versions of the pulses arriving at the receiving end as compared to the original signal. In turn, such delay distortion can make accurate detection of the binary signal very difficult. If the delay distortion characteristics of a channel can be determined in advance, they can be somewhat compensated for at the signal detector. The second type of delay is a delay in the time it takes for the information transmitted (assuming it is accurately reproduced) to arrive at its ultimate destination. For voice conversations, this delay, if long enough, can make voice communications almost impossible. When the round trip delay in a transmitted voice signal reaches on the order of a half second, the receiving and transmitting parties begin to notice this delay in two ways. First, the parties at both ends begin to speak before the other finishes speaking (because each party doesn't accurately know when the other has finished speaking). Second, and more importantly, the parties at both ends begin to hear their own voices echoed back to their ears.

This occurs because the electronics of a typical telephone network cause the far end of almost every call connection to echo back a weakened version of the transmitted signal to the originating end. Although this echo is always present, it is not noticeable until the round trip delay becomes long enough.

All of these deficiencies of a transmission channel have the effect of reducing the channel capacity, or, throughput, of a transmission link. Channel capacity refers to the theoretical upper limit on how much information, in terms of bits per second, can be transmitted through a channel (Wozencraft & Jacobs, 1965).

Circuit-Switched Connections of a Call in a Conventional Telephony Network

VOIP networks must interoperate with conventional circuit-switched networks, in particular the ubiquitous public switched telephone network or PSTN. A VOIP network that cannot interoperate with, and therefore exchange calls with, the PSTN would be of little value to most telephone users. A network's value increases as it can be used to reach more users.

Like other networks, the PSTN uses a hub and spoke architecture of transmission links and switches (Bell Laboratories, 1977, 1983). Each user's telephone is typically connected by an individual circuit (usually referred to as local loop) to a central office switching hub. In a small town, there may be only one central office, whereas in a large major metropolitan area, there could be several dozen central offices.

In each central office hub, there is usually at least one switch. The switches used in traditional telephony are called circuit switches. A circuit switch will route a call over a dedicated path from one transmission link to another for at least the duration of the call. See Figure 4. As will be described more fully below, this contrasts with the packet switching used in the Internet, which does not maintain a dedicated path connecting network links, but instead passes information along shared paths on a packet-by-packet demand basis.

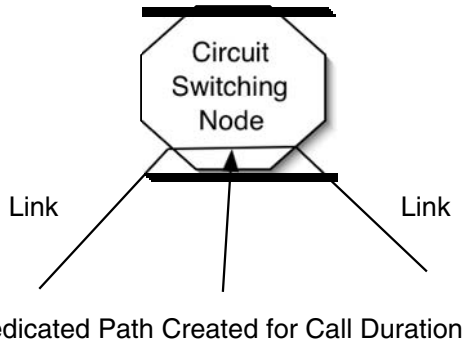


Figure 4: Circuit switched connection between two links.

Originating the Call

As Figure 5 shows, when a caller picks up a telephone to make a call, the local switch in the caller's local central office (Node A) provides a dial tone that the caller hears through the telephone. That dial tone indicates to the caller that the local telephone switch has seized his or her local telephone line and is ready to receive the digits of the telephone number that he or she wishes to dial. The processes of seizing the telephone line and returning dial tone are types of supervisory signaling.

The caller then enters, or dials, the telephone number of the called telephone. If the calling telephone is enabled with touch-tone signaling, the calling telephone signals to the local switch the digits dialed by using different combination of two touch-tones for each digit dialed. Such transmitting of routing information is called address signaling. The caller's local switch Node A detects these tone combinations and, in turn, determines the telephone number the caller is trying to call.

Circuit Switching to the Call's Destination

Each circuit switch has a routing look-up table for determining how to handle a call based on some or all of the calling and called numbers. If the local switch does not have adequate information in its routing tables to make a routing decision, most modern switches will hold the call

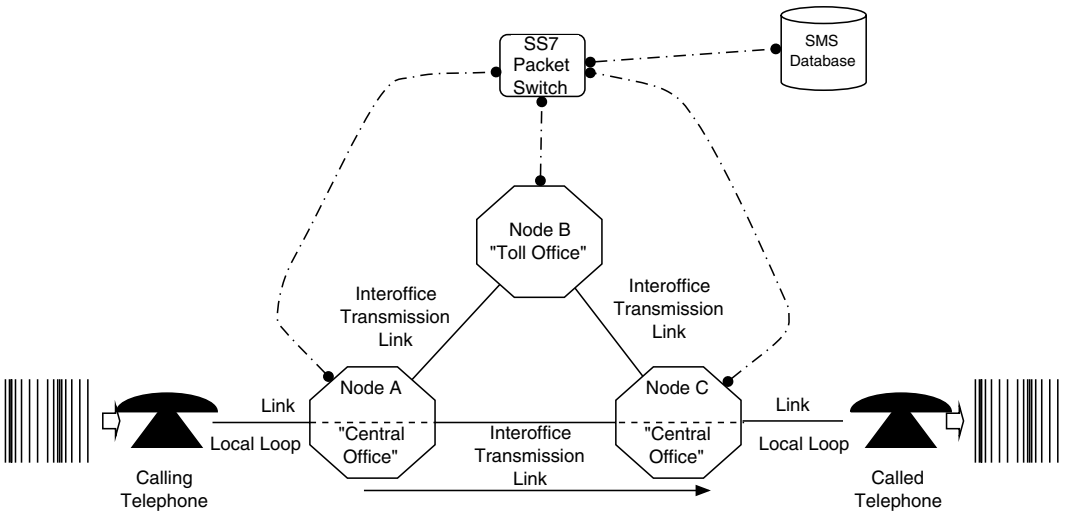


Figure 5: Conventional circuit switched telephony network.

and launch a packet data message to an external database (called a service management system or SMS). That message is routed from the switch to the SMS over a parallel data network called a Signaling System 7 (or SS7) packet network. The SMS returns routing instructions, via the SS7 packet network, to the switch that originated the routing query. The SS7 signaling standard came out of the conventional telephony industry.

If the local switch determines that the call needs to be connected to a local loop attached directly to that local switch, the switch will ring the local loop of the called telephone. At the same time, the switch returns a ringing sound signal to the calling party. On the other hand, if the local switch determines that the called party's telephone is busy, the local switch returns a slow busy signal to the calling party.

As soon as the called party picks up its phone, the local switch creates a circuit-switched dedicated audio communications path between the calling party's local loop and the called party's local loop, thus completing a dedicated audio transmission path between the calling telephone and the called telephone. For the call's duration, that dedicated audio path is available for the transmission of voice signals, in both directions, between the calling and called party's telephones. If one or more of the links used to construct that path are digital, a CODEC in the connecting switch or other connecting device will make the necessary analog-to-digital conversion so that the voice signal may seamlessly traverse the boundaries between the analog and digital links.

If the called telephone's local loop is not directly connected to the calling telephone's local switch, the local switch must establish a connection to the called telephone through another switch or switches. For most modern PSTN switches, this requires that the originating switch launch an SS7 message, over the SS7 network, to reserve a path through other switches in the network for completing the desired communications path to the call's destination. If the far end switch where the called party is located determines that the called party's line is busy, no circuit connections will be made, the path reservation for the connecting links is dropped, and the originating switch will return a slow busy signal to the calling party. However, if the reason the call cannot be completed is network congestion, the originating switch will return a fast busy to the calling party.

If the called party's line is not busy and a complete path to the called party's line can be established, the called party's local switch rings the called party's telephone. While the called party's phone is ringing, the calling party's local switch is returning an audible ringing signal to the calling party. When and if the called party answers, the switches instantly circuit-switch together the links along the reserved path in order to complete a dedicated path between the calling and called telephones.

Finally, it should be noted that some switches communicate interoffice signaling information using non-SS7 interoffice signaling arrangements, such as in-band signaling. Regardless of the supervisory and address signaling arrangement used, the net result of establishing a communications path between the calling and called telephone will be the same.

VOICE OVER INTERNET PROTOCOL

VOIP literally refers to the transmission of a digitized voice signal using digital packets, routed using the Internet protocol or IP. The driving forces for using VOIP are beliefs in its cost savings, flexibility, and the growing desire to combine voice and data transmission on one network. See, e.g., Morris (1998), Cisco VOIP Primer (2002), Matthew (2002).

How VOIP Transmission Works

Because a VOIP call is transmitted digitally, it begins with a digitization process similar to that used in conventional telephony. First, the voice signal is sampled at a rate greater than the Nyquist sampling rate, and those samples are digitized. Whereas conventional circuit switched telephony transmits the digitized samples in a constant stream of synchronized (i.e., equally spaced in time) digital samples, VOIP transmits the digitized voice communications samples in asynchronous (i.e., unequally spaced in time), sequentially numbered packets of data. Each packet (which may contain many voice samples) contains its own IP formatted address information, which allows it to be routed over an IP network. Unlike conventional telephony, where each sample follows the path of the sample before it, each of the IP packets containing several samples of voice data may take an independent path (which is shared with other data packets) to its destination. With each packet potentially taking a different route, the packets often can arrive at their destination out of order (or sometimes not at all). At the far end, the IP and other routing information is stripped from the packet, the voice samples are temporarily collected in a buffer and reordered as required, and then, if all goes well, the original voice signal is reconstructed, albeit, slightly delayed. For VOIP technology to have the functionality and flexibility of conventional telecommunications, some form of call control, in the form of supervisory and address signaling, is required.

Figure 6 shows the basic VOIP transmission scheme.

VOIP Signaling

VOIP signaling refers to signaling that can be used over an IP network to establish a VOIP call. It provides the needed functionality of supervisory and address signaling. VOIP call signaling comes in two predominant competing schemes: H.323 and Session Initiation Protocol (SIP). Each must satisfy the generic needs of telephonic calling. Specifically, each must provide for the following:

The calling phone to address and signal to the called telephone,

The called telephone signaling its availability for receiving calls,

Establishing a transmission path between the calling and called telephone,

The called or calling telephone signaling to the other phone it has hung up, and

The tearing down the transmission path once the call is over.

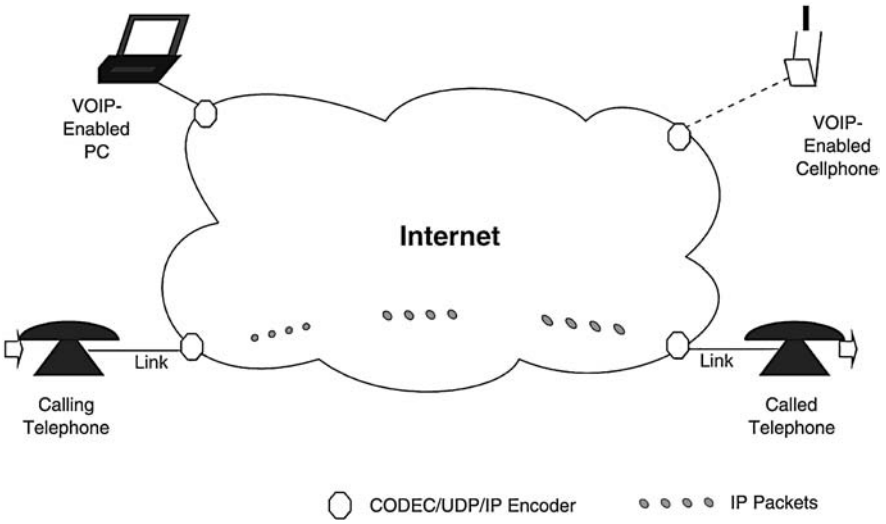


Figure 6: VOIP network.

Figure 7 shows a generic signaling progression for establishing a VOIP call that might occur using H.323-type centralized call control arrangement. First, the calling telephone dials the called telephone's number (1). The calling phone forwards the dialed telephone number (address signaling) to a VOIP call controller—which is a special purpose server. That call controller does a lookup (2) in a database for determining the IP address to reach the called telephone. If the called telephone is on the IP network, a call setup signal is routed to the called telephone to ring the called telephone (4).

While the called telephone is ringing, a ringing signal is sent back to the calling telephone (3), which telephone, in turn, generates a ringing sound in the caller's earpiece. When and if the called party answers the telephone, a series of signals to set up the channel path are returned to the calling and called telephones (5, 6, 7).

A UDP/IP communications path is then set up between the calling and called telephone (8). The digitized voice is transported using the user datagram protocol (or UDP) in the transport layer, with two protocols, RTP (real time protocol) and RTCP (real time control protocol), rather than TCP—which is often used by non-voice data. UDP is connectionless (e.g., packets can take different routes) and can transport data packets without acknowledging their receipt. UDP is nonstop with less address information overhead. The tradeoff of using a UDP path is lower reliability than TCP. UDP packets may be dropped or arrive out of order, but if they do arrive they do with less delay. This is a good tradeoff for voice communications, which is highly tolerant to dropped packets, but relatively intolerant to delay.

RTP over UDP provides packet sequence numbering, so out-of-order and/or missing packets are detectable at

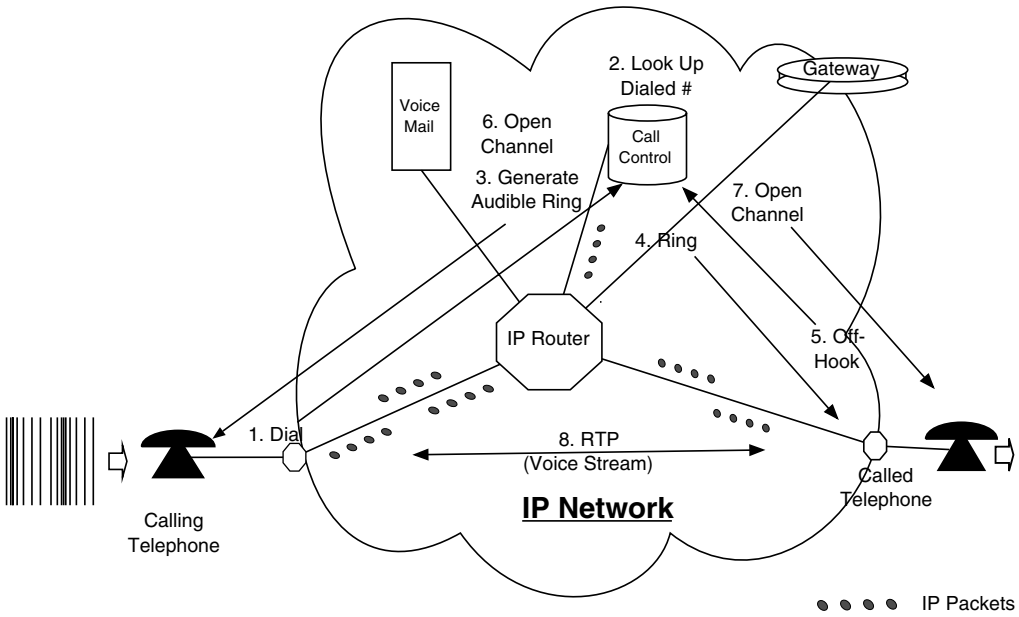


Figure 7: VOIP H.323-type signaling system using central call control.

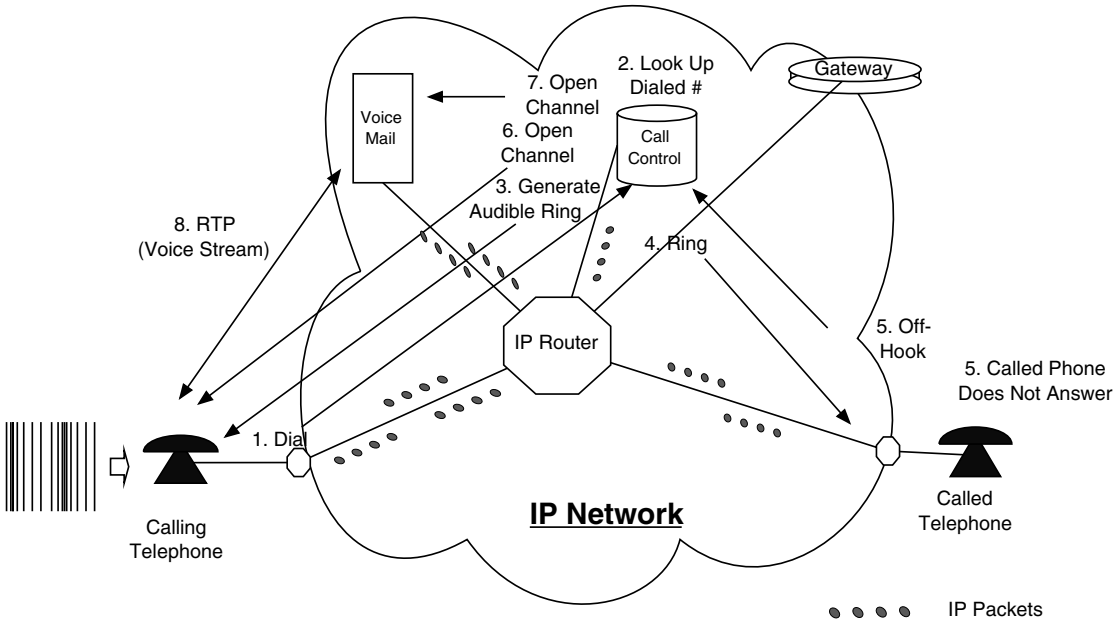


Figure 8: Called telephone unavailable with reroute to voice mail with H.323-type signaling.

the far end. RTCP provides a separate signaling channel between the end devices (e.g., telephones) to allow exchange of information about packet loss, packet jitter, and packet delays, as well as additional information, such as the source's name, e-mail, phone, and identification. (Real-Time Transport Protocol [RTP], 2001, August), and (Streaming Video Over the Internet, 2002).

Additional requirements for a commercial VOIP system include being able to produce information for billing and/or accounting for calls. Similarly, today's users demand that it provide other features, such as Caller ID and voicemail. These capabilities may reside in the call controller, IP telephones, and/or other devices in the IP network. Figure 8 shows how voicemail can be provided with H.232.

H.323 vs. SIP

In addition to H.323, SIP is the major competing standard for VOIP signaling. They have both evolved to offer very similar feature capabilities.

H.323 takes a more telecommunications-oriented approach than SIP. SIP takes an Internet-oriented approach. H.323 is the older of the two (Doron, 2001; Paketizer, 2002). It was developed under the International Telecommunications Union (ITU), a telecommunications standards group, and has gone through various revisions (ITU-T, Recommendation H.323, 2000). The latest version of H.323 standard (H.323 v3) is very robust in that it covers many possible implementations. However, H.323 is considered more difficult to implement than SIP due to its use of binary encoded signaling commands.

H.323 v.3 can be implemented with or without a call control server. Thus, an H.323 v.3 end device (e.g., a telephone) can be designed to either set up a call through a call control server, or set it up directly with another end device without using an intervening call control server. An H.323 v.3 call control server can be set up to relay the communications stream during the call, or the end devices can directly establish the communications RTP streaming channel between themselves. The call control server can be either stateless (i.e., not track a transaction or a session state) or stateful (i.e., track a transaction and/or call session state). The significance of this is that in the stateful configuration, H.323 v.3 is not as scalable. Finally, H.323 v.3 employs signaling protocols that can easily be mapped through a gateway for routing calls between the VOIP network and the public switched network.

SIP is a Web-based architecture that was developed under the Internet Engineering Task Force (IETF). Like URLs and Web e-mail, SIP's messages are in ASCII text format that follow the HTTP programming model, i.e., using a grammar similar to that used to create basic Web pages—resulting in slightly lower efficiency in transmitting signaling information, as compared to the more efficiently encoded binary H.323 signaling messages. Also, SIP is very extensible, leading many vendors to implement variations that may be somewhat incompatible (IETF, SIP, 2003; IETF, SIP RFC2543bis, 2002).

An address used for routing a SIP messages is of the form SIPAddress@xyz.com. As shown in Figure 9, when a phone wishes to originate a call, it transmits an ASCII SIP "INVITE" message (1) addressed to the SIP address of the called phone (e.g., sales@xyz.com., where "xyz.com" is the domain name of the SIP proxy server for the called phone). Using conventional domain name server (DNS) lookup, the internet routes this e-mail-type message to the SIP call control server, which may act as either a proxy server or a redirect server for the domain of the dialed called telephone address (here, xyz.com). In order to determine where the telephone is located, the proxy server or redirect server will query a location server (2), which will return the routing directions. What happens next depends upon whether the call controller is acting as a proxy server or a redirect server for the called telephone.

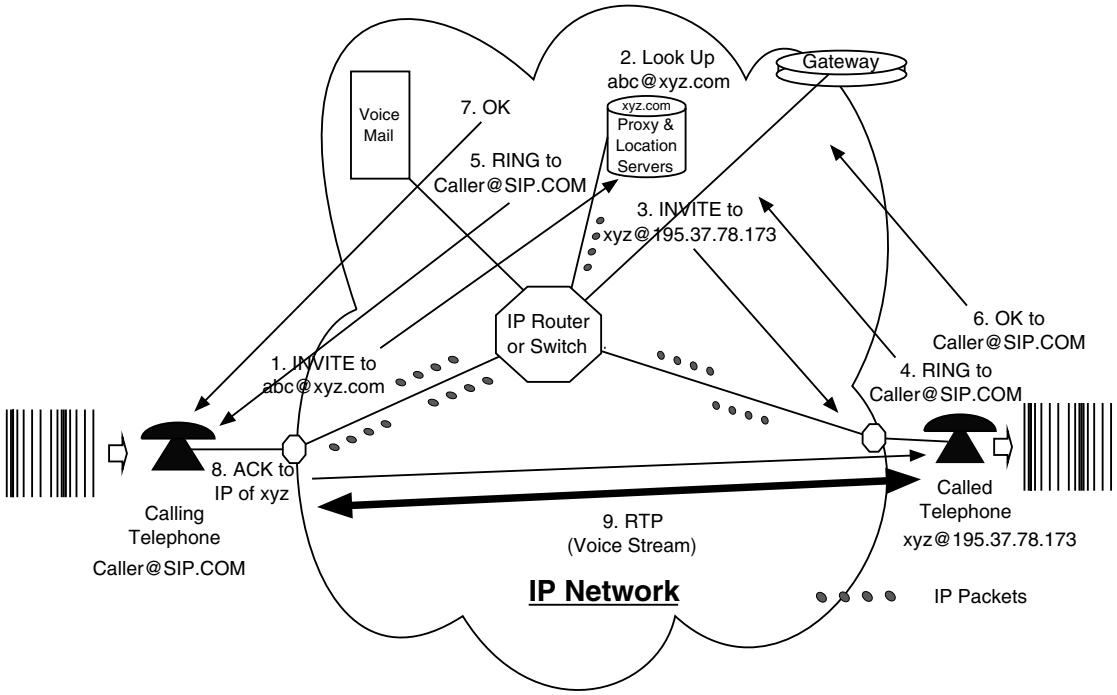


Figure 9: VOIP call routed using SIP signaling.

For a proxy server, the location server will return the IP address for the called telephone. Using that IP address, the SIP INVITE message will be directed to the called telephone (3) along with the calling telephone's IP address. The called telephone will then return a "RING" message to the proxy server (4), which then forwards that RING message to the originating telephone (5). When the called phone answers, an OK message is returned (6, 7), which includes the called telephone's IP address. Finally, the originating telephone (which now knows the IP address of the destination phone from the INVITE/OK message exchange) sends an "ACK" message to the IP address of the called telephone (8). The originating telephone then establishes an RTP communications link with the called telephone (9).

For a redirect server, the location server will return the redirected (i.e., forwarding) e-mail-style SIP address of the called telephone (e.g., sales@home.com, or joe@home.com). The redirect server will then forward the INVITE request to the proxy or redirect server associated with that redirected address, and then the steps enumerated above will take place.

Each location server must track the IP address of each of the telephones in its SIP domain. Thus, each SIP telephone must register with its domain's location server via a registration server of its telephone service provider. An individual telephone can be registered with any registration server with which the user has a service arrangement. Registration binds each SIP telephone's IP address to its SIP address in its service provider's domain.

Note that a SIP call control server's primary purpose is to handle the routing of initial supervisory and address signaling information. Also note that, after the initial exchange of supervisory and address information, SIP end devices establish and maintain the communications channel without involvement of the SIP call control server. Like H.323 v.3, the SIP packets that carry the signaling messages almost always follow a different path from the path taken by the communications data. Finally, with SIP most of the intelligence resides in the end devices, as compared to being in the network, as is the case with conventional telephone networks and with H.323.

Because SIP proxy and redirect servers typically do not track a call's status after the call is set up, SIP is often viewed as being more scalable than H.323. When a call control server is configured to track call status, its resources must bear the added burden of such monitoring.

SIP's ASCII encoding is considered more extensible and open than the binary encoded signaling of H.323 v.3. SIP uses a very generic syntax for messages, which can be customized to fit the needs of different applications resident on end devices. For analysis of the various (and somewhat controversial) comparisons of SIP and H.323, see Dalgica and Fang (1999).

Integrating VOIP Into Conventional Circuit-Switched Telephony Networks

As previously noted, the value of a telephony network is a direct function of how many telephones are directly or indirectly connected to it; thus a VOIP network must be able to exchange calls (i.e., internetwork) with the PSTN such that a VOIP user can originate telephone calls to and receive telephone calls from a telephone user who is connected to the PSTN. In addition, for conventional telephone providers to deploy VOIP technology inside their networks, VOIP technology's presence must be imperceptible to their existing base of telephone users.

Figure 10 illustrates how a VOIP call can make a connection from a phone connected to an IP network to

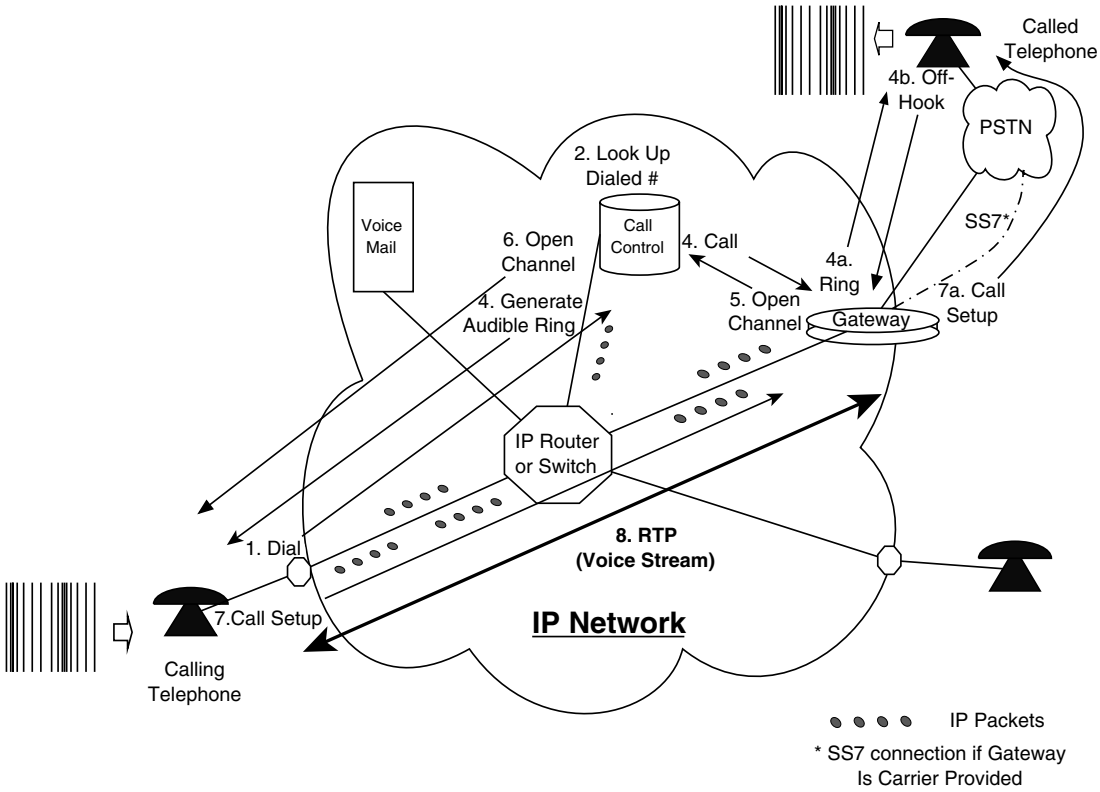


Figure 10: VOIP signaling internetworking with public switched network (PSTN).

a phone on a PSTN network. A device, called a gateway, is used to translate signaling messages across the VOIP/PSTN network boundary and to transform the jittery voice packets on the IP side of the gateway into a synchronous stream of voice data information (if there is a digital voice circuit on the other PSTN side of the gateway) or an analog voice signal (if there is an analog voice circuit on the PSTN side of the gateway).

On the IP side, the VOIP signaling to the gateway looks much the same as the signaling that would be done to another end device on the IP network.

ENUM, the Fully Interoperable Numbering Plan

ENUM is a new standard for numbering plans that would allow seamless telephone number addressing between conventional and VOIP telephony. See, e.g., Neustar (2003) and IETF, Telephone Number Mapping (ENUM) (2003). It unifies Internet and conventional telephone addressing schemes by mapping E.164 (i.e., conventional telephone) numbers to a URL Internet (and SIP)-friendly format. With ENUM, a single global digital identifier system can serve equally subscribers attached to the PSTN or the Internet. The same identifier can be used to identify multiple devices, such as plain telephones, fax, voicemail, and data services, regardless of whether they are on the PSTN or public Internet, thus conserving scarce E.164 numbers and domain name resources.

As an example, a telephone number 1-305-599-1234 would map onto the URL 4.3.2.1.9.9.5.5.0.3.1.E164.arpa. A DNS query on this domain name would return a number of records, each listing a specific service registered to the owner of the E.164 number. Devices and services attached anywhere on the public Internet or PSTN could be registered to this universal, fully portable, single number identifier of their registered owner, allowing mobility around both the PSTN and the public Internet.

The ENUM standards raise new issues of privacy, security, and administration. Also, final agreement between Internet and traditional telecom industry represented by the ITU is still pending.

Quality of Service Issues

Transmission and routing can introduce effects that degrade the quality of VOIP. VOIP signals are not immune to the deficiencies of the facilities that transport and route them. These deficiencies can cause packet loss, jitter, and delay.

Packet loss refers to the loss of packets containing some of the voice samples during transmission. The loss might be caused by high bit error rates in the UDP transmission channel, misrouted packets, and/or congestion causing intermediate routing devices to drop packets. Because voice transmission is very sensitive to delay, the reassembly of the voice signal at the receiving end of the call cannot typically wait for the retransmission of erroneous or misrouted packets. However, voice transmission has a high tolerance for packet loss—mainly because the ultimate receiver of the signal (the human ear and/or intervening CODEC) does a good job of interpolating (i.e., filling in the gap) where a packet has been lost. There is a limit to how much packet loss can be tolerated. That limit

depends upon several factors, including: (1) the nature of the sound being transmitted, (2) the correlation in time between the packets lost (i.e., are they bunched together or widely dispersed in time), and (3) the randomness of the losses. Because they do such an efficient job of squeezing out redundancy, some CODECs that use data compression algorithms are not able to recover from the loss of as few as two packets in a row.

Jitter can also contribute to lower quality. With IP traffic often taking multiple routes and/or mixed in with bursty IP packet traffic, the interarrival times of the packets at the receiving end may be irregular and/or the packets arrive out of order. Jitter can be overcome by buffering the packets—i.e., temporarily storing them long enough to reorder them before forwarding them to the decoder. However, buffering has the negative consequence of adding more delay.

Delay in the transmission of voice packets can be minimized in several ways. One is overbuilding the IP network, i.e., ensuring that there is always excess capacity in the IP network for all traffic, including during peak traffic periods. Alternatively, priority routing can be afforded to voice traffic, at the expense of lower priority data traffic—which is more tolerant of delays. Separate treatment of voice from data can be done by virtual segregation or physical segregation from lower priority traffic. The former can be done at either the network or data link layer. For example, segregation can be done at the data link layer by using separate ATM channels for voice and data traffic and then assigning ATM-based priority treatment to the ATM channels carrying voice traffic. Where voice and data are mixed on the IP network, identifiers can be used to indicate to the intermediate network components (such as routers) that designated traffic (such as voice traffic) should be given priority.

An example of this last method is RSVP, the Resource Reservation Protocol. RSVP allows a VOIP application to request end-to-end QoS guarantee from a network (Cisco, VoIP Call Admission Control Using RSVP, 2003). If the guarantee cannot be provided, the call will not be allowed to go through. Where the guarantee cannot be secured, the traffic might be redirected to an alternate network or blocked (resulting in VOIP users receiving an "equipment busy" signal). At the current time, priority schemes such RSVP typically do not work over the public Internet (with a large "I"). This is because, among other reasons, the economic incentives are not there for intermediate Internet providers to honor any type of prioritization routing scheme, given that their reimbursement is the same for all traffic—regardless of its priority designation. Therefore the current market structure for public Internet backbone routing prevents the realization of a higher quality of service for VOIP traffic over the public Internet.

To differentiate themselves, some Internet backbone providers are introducing prioritization schemes such as MPLS-based networks, which are ATM-like in their attributes, but operate at a mixture of layer 2 and 3 protocols in pure IP environments. As competition intensifies, public networks are expected to become friendlier to real-time services such as VOIP. Quality of service is discussed in detail elsewhere.

The Costs and Savings of Using VOIP

One source of VOIP's cost savings over conventional telephony is its ability to employ transmission more efficiently due to both the extensive use of compression algorithms and the statistical nature of its information transmission. However, such efficiencies will tend to be most significant to private networks and/or network providers whose transmission networks are capacity-constrained.

A second source of VOIP's cost savings is lower capital cost per call using lower cost switching devices (i.e., Internet routers and switches). Again, networks with sunk investments in conventional technology with excess capacity would derive little benefit from such savings.

A third source of VOIP's savings is lower costs of administration, particularly in enterprise environments. A good deal of administrative cost is incurred in enterprises to accommodate the movement of telephone users within the enterprise. Each time a user moves to another office or enters or leaves the firm, the routing tables and directory of a conventional telephone system must be updated, often manually. VOIP's self-registration feature eliminates these administrative costs.

A fourth source of VOIP's savings comes from the "economies of scope" that VOIP can achieve by its ability to intermingle with other traffic on data networks, eliminating the need to segregate voice and data traffic, as is often done with conventional telephony. These savings are most easily exploited in the LAN and WAN enterprise environment or by data-centric carriers who wish to combine their voice and data traffic. Administrative savings also come from eliminating the conventional regime of separate administrative staffs for voice and data.

Finally, VOIP traffic on a data network looks like any other data on that network. This allows some carriers and enterprise users to avoid some of the economically distorting taxes that local, federal, and foreign regulatory regimes place on pure voice traffic, but not on data traffic. It is important to remember that the realization of these savings are application-specific and may not be realized in every situation. See, e.g., Morris (1998) and "Cisco Seeks Bigger Role in Phone Networks" (March 15, 2003).

Security Issues for VOIP

Security is a concern with VOIP, particularly because of the distributed nature of the call control, much of which is handled between end devices (Cisco, SAFE: IP Telephony Security in Depth, 2002). Some of the things that make VOIP attractive, e.g., self-registration of end devices and software-controlled PCs acting as end devices, are also the sources of VOIP's vulnerability. Security issues include four categories: (1) eavesdropping, (2) toll fraud, (3) identity spoofing, and (4) IP spoofing.

Eavesdropping refers to an unauthorized party "listening" to the packets and, in turn, being able to listen to the voice conversation. This problem exists with both VOIP and conventional analog/digital voice communications. In both cases, the simplest method of preventing this problem is to encrypt the digital signal at its source, the originating telephone. The problem encryption brings is overhead and computational load, which can introduce delay. Also, encryption can create problems

for law enforcement agencies where they have a warrant to wiretap a telephone conversation.

Toll fraud is the problem of unauthorized persons making telephone calls over the network. This can be caused by unauthorized users originating calls on the IP network either on legitimate phones or over illegitimate phones. As with conventional telephony, this can be combated with security codes and other authentication methods.

Identity spoofing refers to a hacker tricking a remote user into believing he or she are talking to the person he or she dialed on the IP network, when, in fact, he or she is talking to the hacker. Again, security codes and other authorization methods are helpful here.

IP spoofing refers to theft of IP identity, where one end device is able to convince the IP network that its IP address is the same as a legitimate device's IP address. This allows the device with the fraudulent IP address to intercept calls and/or perform toll fraud using that IP address. Spoofing the IP address of a gateway allows eavesdropping on telephone calls.

Some generally accepted recommendations for minimizing many of these security problems are to disable self-registration of VOIP end devices after initial network installation, to segregate voice from data traffic at level 2 or 3, and to use a stateful firewall at the PSTN gateway. As noted above, segregating data from voice services also provides the added benefit of maintaining different quality of service for data and voice.

CONCLUSION

VOIP holds great promise where the convergence of data and voice can occur. Internetworking and overcoming QOS issues remain some of the biggest challenges.

GLOSSARY

Analog signal A continuous signal that, at any point in time, can have an infinite number of possible values and that is typically analogous in some characteristic to another signal or physical phenomenon.

Asynchronous transfer mode (ATM) A network transfer method, employed at the data link layer (Level 2), for high-speed switched packet routing, which can establish virtual channels with a specified QoS.

Channel capacity The theoretical upper rate limit, in bits per second, at which information can be transmitted over a transmission channel.

Circuit switch A switch that makes a temporary or permanent dedicated transmission path between two transmission links that are attached to that switch, based on signaling information received prior to establishing the dedicated path.

Digital encoding Encoding a signal in the form of a string of 1s and 0s.

Digital transmission Transmission of information encoded as 1s and 0s.

Internet A global public network based on the "Internet protocol," connecting millions of hosts worldwide, and for which users often pay a flat fee to access, with little or no charge for transmitting each packet of information. (Outside the U.S., Internet access is often measured and charged on a usage basis, e.g., minutes or units of data.)

Internet protocol (small "i") or IP A packet switching protocol used for routing packets over and between and private networks that is "connectionless" (i.e., each packet making up the same message may take a different route to reach the ultimate destination).

Node A point of connection between transmission links, which may contain switches, and/or may contain converters to interconnect transmission links with differing modalities (e.g., for connecting wire links to wireless links, non-digitally encoded links to digitally encoded links, or fiber links to copper wire links).

Packet router A type of packet switching device that typically routes based on Level 3 network address information (such as an IP address) and typically has the ability to choose optimal routing based on dynamically changing criteria and routing tables.

Packet switch A type of packet switching device that routes packets of data between links based on address information associated with each packet. A Level 3 switch uses network addresses, such as IP addresses, to route packets of data. Level 2 switches uses data link layer addresses (which are typically local and/or hard-coded) for routing.

Public switched telephone network (PSTN) A circuit-switched network that is provided by regulated common carriers who offer their voice telephone services to the general public.

Quality of Service (QoS) A set of performance parameters or criteria, such a bandwidth, jitter, packet loss, and delay, prespecified for a service.

Transmission The movement of information (whether or not digitally encoded) from one point to another via a signal carried over a physical medium, such as wires, fiber, radio, or light.

Transducer A device actuated by signal power from one system and supplying signal power in another form to a second system (e.g., a telephone receiver earpiece actuated by electric power of a received transmission signal and supplying acoustic signal power to the surrounding air, which the telephone user can hear, or a telephone microphone that has a quartz crystal that produces electrical signal power for transmission over wires from the mechanical acoustic power originating from the telephone user's voice).

Transmission link A transmission path connecting two nodes.

Voice communication The transmission of information contained in a voice signal.

CROSS REFERENCES

See *Circuit, Message, and Packet Switching; Digital Communication; Internet Literacy; Public Networks; TCP/IP Suite; Web Quality of Service; Wide Area and Metropolitan Area Networks.*

REFERENCES

ASCII Table. Retrieved March 23, 2003, from http://web.cs.mun.ca/~michael/c/ascii-table.html

Bell Laboratories (1977, 1983). *Engineering and operations in the Bell System*. Murray Hill, NJ: Bell Laboratories.

Castelli, M. (2002). *Network consultants handbook*. Indianapolis, IN: Cisco Press.

Cisco, VoIP call admission control using RSVP (2003). Retrieved March 25, 2003, from http://www.cisco.com/univercd/cc/td/doc/product/software/ios121/121newft/121t/121t5/dt4trsvp.htm

Cisco, VOIP primer (2002). Retrieved March 25, 2003, from http://www.cisco.com/univercd/cc/td/doc/product/access/acs_mod/1700/1751/1751 swg/intro.htm

Cisco, Understanding codecs: Complexity, hardware support, MOS, and negotiation (2002). Retrieved March 25, 2003, from http://www.cisco.com/warp/public/788/voip/codec_complexity.pdf

Cisco, SAFE: IP telephony security in depth (2002). Retrieved March 25, 2003, from http://www.cisco.com/warp/public/cc/so/cuso/epso/sqfr/safip_wp.htm

Dalgic, I., & Fang, H. (1999). Comparison of H.323 and SIP for IP telephony signaling. Retrieved March 25, 2003, from http://216.239.57.100/cobrand_univ?q=cache:PNnMM0MUWcsC:www.cs.columbia.edu/~hgs/papers/others/Dalg9909_Comparison.pdf+dalgic&hl=en&ie=UTF-8

Doron, E. (2001). SIP and H.323 for voice/video over IP—Complement, don't compete! *Internet Telephony*. Retrieved March 25, 2003, from http://www.tmcnet.com/it/0801/0801radv.htm

Internet Engineering Task Force (IETF) SIP (2003). Retrived March 25, 2003, from http://www.ietf.org/html.charters/sip-charter.html

Internet Engineering Task Force (IETF) Telephone number mapping (ENUM) (2003). Retrieved March 25, 2003, from http://www.ietf.org/html.charters/enum-charter.html

Internet Engineering Task Force (IETF), SIP RFC2543bis (2002). SIP: Session initiation protocol, SIP WG Internet Draft. Retrieved March 25, 2003, from http://www.jdrosen.net/sip_bis.html

ITU-T, Recommendation H.323 (2000). Retrieved March 25, 2003, from http://www.itu.int/rec/recommendation.asp?type=items&lang=E&parent=T-REC-H.323-200011-I

Morris, R. L. (1998). Voice over IP telephony: Sizzle or steak? Retrieved March 25, 2003, from http://members.aol.com/_ht_a/roym11/LoopCo/VOIP.html

Neustar (2003). Retrieved March 25, 2003, from http://www. enum.org

Newton, H. (1998). *Newton's telecom dictionary* (14th expanded ed.). New York: Flatiron Publishing.

Cisco seeks bigger role in phone networks (2003, March 3). *New York Times*. Retrieved March 25, 2003, from http://www.nytimes.com/2003/03/03/technology/03CISC.html?tntemail1=&pagewanted=print&position=top

Packetizer, T. M. (2002). Comparisons between H.323 and SIP. Retrieved March 25, 2003, from http://www.packetizer.com/iptel/h323_vs_sip/complist.html

Real-Time Transport Protocol (RTP) (2001, August). Retrieved March 25, 2003, from http://www.cs.columbia.edu/~hgs/teaching/ais/slides/rtp.pdf

Sanford (1999). Packet voice technology: Cheap talk? Retrieved March 25, 2003, from http://www.applied-research.com/articles/99/ARTicle10Sanford.htm

Truxal, J. G. (1990). *The age of electronic messages*. Cambridge, MA: MIT Press.

Streaming video over the Internet (2002). Retrieved March 25, 2003, from http://www.streamdemon.co.uk/tranproto.html

Wozencraft, R., & Jacobs, M. (1965). *Principles of communications theory*. New York: Wiley.

W

Web-Based Training

Patrick J. Fahy, *Athabasca University*

WEB-BASED TRAINING (WBT): BACKGROUND

As part of corporate health, even survival, companies and training institutions globally have recognized the need to provide relevant and flexible training. Professional development (PD) in the form of upgrading, re-training, and various educational opportunities is seen as enhancing the skills of valued employees, helping organizations maintain their competitive advantage by developing (and thereby retaining) experienced people.

Well-designed Web-based training (WBT) can offer valuable advantages over other types of training delivery in a wide variety of public and private environments: training time and travel can be reduced, even eliminated, lowering costs; materials stored on central servers can be continually revised and updated, assuring currency and enhancing quality; training content is more consistent, supporting higher standards; greater efficiency (chiefly the result of individualization) can increase trainee learning and satisfaction, improving motivation; and production and delivery of training programs may be more systematic, improving the cost-effectiveness of development.

At the same time, using the World Wide Web (WWW) for training presents some challenges: existing training materials must usually be redesigned, sometimes extensively; bandwidth limitations (often at the user's end, in the "last five feet" of the communications chain) may restrict or even prohibit use of multimedia by some trainees; all participants (trainees and instructors) must learn new skills to use WBT effectively; and an initial investment (sometimes substantial) in equipment and expertise may be needed. Other factors in the structure or culture of a training organization may also need to change to make WBT feasible. Dropout rates (admittedly often a problem in WBT) may indicate the health of WBT programs: high rates may mean a mismatch between trainees' expectations and the instructional design of the training material, or may reveal a lack of leadership or management support (Frankola, 2001).

In this chapter, WBT will be discussed from theoretical and practical perspectives: important training principles are reviewed briefly, including basic concepts now common in WBT; practical problems in WBT are considered, as well as the strengths and weaknesses of this mode of training delivery; and finally the prospects for the future of WBT, and some of the pedagogic, technical, and economic assumptions on which the optimistic predictions depend, are considered.

Training Principles and Technological Developments Supporting WBT

Pioneering Ideas in Training

In the first half of the 20th century, pioneering researchers such as Thorndike (1971), Dewey (1938), Skinner (1971), and Keller (1968) conducted research that began identifying fundamental learning principles. (While these figures wrote and researched in the fields of psychology and education, their theories have evolved so that they are now used in the design of effective teaching and training of all kinds, including WBT.) Thorndike's three fundamental behavioral laws were among the first discoveries: (1) repetition strengthens any new behavior; (2) pleasure or reward associated with a particular behavior increases the likelihood the behavior will be repeated, while pain or lack of reward may diminish the likelihood; and (3) an individual's personal readiness is crucial to the performance of any new skill or behavior (Saettler, 1990).

Dewey added that individual trainee differences were crucial in the success of training. Dewey and Piaget (1952) both recognized the importance of each individual learner's personal background, and advocated that trainees' experiences and previous learning be considered

carefully and accommodated as possible in any new training situation.

Skinner applied and extended the behavioral principles emerging from the new science of psychology, launching the "teaching machines" movement of the 1950s and 1960s. By describing the teacher or trainer as "the manager of the contingencies of reinforcement" in the learning process, Skinner helped found the fields of educational technology and information processing psychology. Besides a new role for the instructor, Skinner's work illustrated points vital to the subsequent development of technology-based training, including the value of "the program over the hardware," and the critical importance of the learning materials and the organization of the learning environment (Saettler, 1990).

Keller's Individually Programmed Instruction (IPI) model applied Skinner's discoveries about the importance of instructional design. The IPI model (often called the *Keller Plan*) emphasized individual differences in instruction and evaluation. IPI stressed such principles, later core to WBT, as self-pacing, mastery before advancement, high-quality materials, tutoring help, prompt feedback, and practice testing (Fox, 2002). Experiments in teaching Morse code to World War II recruits demonstrated convincingly that these principles could dramatically increase the efficiency of training.

Innovations such as IPI were impressive, but they encountered resistance for several significant reasons relevant to WBT:

- As innovations, technology-based training models often make new demands upon institutions, trainers, and trainees. In particular, technology-based training delivery increases the responsibility of trainees for their own learning, while reducing instructors' "platform behavior," serious flaws in the eyes of some trainers (and some trainees).
- Technological innovations may initially result in a judgment of "no significant difference" in performance when evaluated. Time may be needed to reveal their true value.
- Tepid managerial support may doom a technological innovation, especially if time and funds are needed to prove its actual potential.
- Individualized training models usually require more advance planning and preparation, while (at least initially) increasing workloads on instructors and administrators, and possibly destabilizing programs while adjustments are occurring. Participants must be aware of these, and must be prepared to work through them.
- When granted responsibility for managing their own learning, some trainees may respond with demands for more individual treatment, including access to records of personal achievement, and remedial and accelerated options. Overall, in individualized programs trainees expect their individual performance, needs, and preferences to be acknowledged.

In the 1960s and 1970s, developments in instructional design, cognitive and behavioral psychology, and organizational analysis converged to produce new tools and approaches for the design and delivery of training, which would quickly impact the emerging field of technology-based training, and, eventually, WBT. Key figures whose work influenced the practice of training (individually, in collaboration, and collectively), in addition to those already mentioned, included Robert Mager (1975), Patrick Suppes (1978), Robert Glaser (1978), Robert Gagne and Leslie Briggs (1979), Leslie Briggs and Walter Wager (1981), and Walter Dick and Lou Carey (1978). There were certainly others, but these individuals led the way.

The Internet as a Training Platform

The earliest forms of the Internet emerged as training was being transformed by a new understanding of learning itself. The fact that the Internet today supports a staggering array of commercial and educational enterprises and utilities (browsers, search engines and indexes, media players and plug-ins, mark-up languages and authoring tools, etc.) is due in part to the commitment to openness and accessibility, and recognition of the importance of interaction as a learning support, made in the early days of its development. The validity of the vision of the early developers of the Internet of an infrastructure of open, flexible protocols and common standards can be seen in the fact that the Internet's first "killer app" was electronic mail (e-mail), still the most used tool on the Internet today. The early developers' commitment to cooperation and collaboration reflected in the modern Web's friendliness and durability helps make it a powerful tool for (among myriad other things) delivering accessible training (Leiner et al., 2002).

High-Technology and Training

Origins: Computer-Based Training (CBT)

As programmed instruction and teaching machines declined in the late 1960s, and before the early Internet emerged from the developers' labs, the first commercial computer-based training (CBT) systems appeared. Initially, CBT systems such as PLATO (*Programmed Logic for Automated Teaching Operations*) and TICCIT (*Time-shared, Interactive, Computer-Controlled Information Television*) were costly, experimental, and rudimentary: text was the mode of presentation, using only occasional simple line drawings or diagrams, with minimal or no animation, sound, or color.

Technical milestones passed quickly in CBT over the next two decades: IBM mainframes were programmed to teach binary arithmetic; mainframe PLATO and TICCIT were used in college teaching (TICCIT was later adopted by the U.S. Navy); authoring languages such as *Coursewriter* and *PILOT* permitted instructors to produce instructional and testing materials for mainframe delivery and, later, for PCs (Rutherford, Patrick, Prindle, & Donaldson, 1997); and multimedia platforms using laserdisc and, later, CD-ROM were perfected for storage and portability.

The personal computer revolutionized and gave a huge boost to CBT. Apple, IBM, and IBM "clone" personal computers (PCs) became increasingly powerful. As prices dropped, the appeal of desktop technology grew; as authoring capabilities increased, enthusiastic individuals spent money, time, and energy on CBT programming projects (sometimes duplicating the efforts of others, and,

due to the lack of careful instructional design, frequently producing materials of marginal training value). The need was clear for a way to link desktop computers for delivery of CBT, in order to increase access to quality information and training resources, reduce needless duplication and unnecessary competition, and facilitate the fundamental goals of collaboration and cooperation among developers and users.

Computer-Based Communications

While specialized uses of proprietary CBT systems were expanding and the lack of effective communications among users was creating obvious training inefficiencies, promising first steps toward use of computers in an open, publicly accessible network were being made, initially, in 1962, with the U.S. military's ARPANET initiative. ARPA (Advanced Research Project Agency) proposed an online research network to link defense contractors and academics with the military, and with each other. Attempting to use the computers of the time as interactive communications devices to promote collaboration was considered novel, and was questioned technically as well as practically by those who still saw the computer as only "an arithmetic engine" (Hauben, 1994). However, in 1972, by the time the final ARPANET report was produced, the feasibility and usefulness of computer-based networked communications was established. Legacies of ARPANET, including tools such as file transfer protocol (FTP) and TCP/IP (*transfer control protocol/Internet protocol,* a reliable packet-addressing, flow-control, and loss-recovery tool), were available and were subsequently incorporated in the first Web browser, *Gofer,* in the early 1990s, and in the first true graphical browser, *Mosaic,* in 1993. Even more importantly, the principles of free communication and interaction were also reflected in the open and unregulated architecture of the early commercial Internet.

USING THE WEB FOR TRAINING

Strengths

The early Internet, with its ease of access and openness, appealed to trainers who were previously forced to rely upon standalone CBT, or pay high costs to access proprietary online training networks. However, previous experience with CBT had shown that a networked training system would need to include certain features to be maximally effective, such as a common interface and delivery format, interoperability among different desktop and server systems, and aids to communication and collaboration, to make development and production processes more efficient, in the form of flexible and robust interactive capabilities. The Web, because of its fundamental openness and flexibility, from the beginning proved impressively capable of accommodating these and many other functions.

A Common Training Platform

The Web had evolved to commercial viability by 1995. (In August of that year, *Netscape* went public with a hugely successful IPO on NASDAQ.) It was obvious that for the Web's widespread adoption, the PC would be the key technology: when networked, the computer would be the critical convergence device for this new means of communication and collaboration. With access to the Internet via a properly equipped computer, anyone, including training providers and their clients, could share the Web's growing "super-network" of services and resources, including training, from virtually anywhere in the world.

Interoperability

Previously, the costs and complexity of developing online materials were greatly inflated by the need to produce versions meeting the specific requirements of hardware platforms, configurations, and operating systems. The Web, using cross-platform "mark-up" (as opposed to programming) languages such as HTML (*hypertext markup language*), SGML (*standard general markup language*), and the powerful XML (*extensible markup language*), made hardware and operating system characteristics largely irrelevant to Web access.

Although markup languages and the Web's interoperability capabilities eliminated many barriers to access arising from system differences, issues still persist today because the Web grants users great freedom in their choices and configurations of software and hardware. Issues arise from, for example, differences in display hardware (monitors) and screen resolutions, Internet connection speeds (still considered the greatest limitation to access of new, multimedia material), and user settings for fonts, background colors, display resolutions, etc., which can dramatically change the appearance and effectiveness of training materials (Jones & Farquhar, 1997). Problems may also arise with new versions or updates of common browsers, plug-ins, and other required software, which occasionally fail to perform as well as previous versions, and may even contain fatal bugs. To address these potential problems, training programs often publish suggested standards for hardware and software (including version or "build" numbers and any required patches or service releases for software), to guide trainees in upgrading their systems for full compatibility.

Training Efficiency

Though costs of initial development were high, it was soon clear that properly designed WBT could deliver impressive results, especially for the highly motivated and wherever cutbacks threatened to affect the quality of face-to-face programs (due to larger training group sizes, limited modes of content presentation, declining opportunities for remediation or individual tutoring, reduced interaction generally, etc). Some advantages favoring quality WBT over typical site-based face-to-face delivery included self-pacing and individualization; greater emphasis on learning, less on instruction (seen in the emphasis on high-quality materials); learner control and autonomy in the training process; more flexible and convenient remote access to training opportunities; quicker and more personalized feedback; peer-to-peer social interaction; and timely, on-demand access to preparation, remediation, review, and outside resources. These advantages were observed early, and led some observers to claim that WBT could be superior in quality to even well-designed and properly conducted face-to-face training, especially when training group sizes grew larger (Harapniuk, Montgomerie, & Torgerson, 1998; Kaye, 1989; Wagner, 1994).

Costs and Convenience

Employers soon found that both direct (travel and subsistence) and opportunity costs (employee fatigue and lost productivity, time away from the workplace, stress) could be reduced by WBT, especially if the training package was designed to take full advantage of the Web's potential efficiencies and conveniences. As bandwidth increased, a progressively more "media-rich" experience could potentially be delivered to trainees almost anywhere. (Ironically, although asynchronous training provided maximum convenience and lower direct delivery costs, the temptation to use rich multimedia-based synchronous [same-time] sessions, more feasible as bandwidth increased and became cheaper, created some of the same problems of inflexibility and inconvenience as face-to-face training formerly did for users who lacked access to high-bandwidth Internet services.)

Weaknesses

Some basic weaknesses were soon evident in Web-based training, too: unless developers employed instructional design principles skillfully and conscientiously, the quality and integrity of materials were sometimes uneven; navigation among linked sites could become a nightmare for the unwary or inexperienced; and users needed to possess and consistently practice self-discipline to avoid being side-tracked by online distractions.

Quality

Quality continues to be a problem on the Web, because the public Internet by definition is a loosely coordinated (not controlled) network of networks. No one is accountable for the quality of what is found anywhere on the Web, or for maintaining functionality: anyone may post information (or misinformation), so points of view masquerading as fact are common (Warren, Brunner, Maier, & Barnett, 1996), and important links may simply fail to connect if not updated frequently. For these reasons, the provenance of any Internet-based information must always be questioned, and linking to Internet materials for training is risky as there is nothing to prevent content from changing, or links from disappearing altogether. Increased security of access and control over content are arguments for *intranet* delivery of training.

Structure and Navigation

Just as no one oversees the quality of materials on the Web, no one assures that Web pages articulate clearly. Retracing one's steps can be a challenge, even for experienced users. The Internet is like a library where all the books have lost their identification codes, and some their bindings. The unwary may quickly become lost in cyberspace.

User Control and Orientation

A strength of the Web—the freedom to explore freely—can be a major weakness for some. The lack of restrictions and navigation guides forces users to make their own choices from a huge number of Web offerings; the inexperienced, immature, impulsive, easily distracted, or learning disabled may be challenged, even overwhelmed. At the same time, the opportunity to seek out related but peripheral information, to pursue an interesting detail, and to dig deeper into a subject are all celebrated differences between Web-based and traditional learning, for those with the discipline and skills needed to use them wisely. For these reasons, proper orientation and training of users is a prerequisite to successful WBT.

WBT'S CHALLENGES

Web-based training, as an innovation, presents challenges: different roles and responsibilities for instructors and trainees; changes in teaching methods; the importance of provision for individual differences; and new financial and economic realities.

New Roles

Web-based training creates changes in the ways trainees connect with the trainer, the content, and the learning system. As described below, because WBT allows shifting of place and pace of learning, roles change; the focus is on the trainees' skill development, and the tutor or trainer consequently becomes less the "sage on the stage" and more a "guide on the side" (Burge & Roberts, 1993). Similarly, if permitted, equipped, and disposed to do so, trainees may assume more responsibility for their own learning, including accessing outside materials and communicating as needed with the trainer and with other trainees. Overall, successful WBT changes trainees' experiences, providing greater individualization, making feasible conveniences such as self-pacing, on-demand review, acceleration, and practice testing, and providing ready linkages to other people and resources.

As the early pioneers of technology-based training found, the quality and completeness of the learning materials are critically important, as are the appeal, intuitiveness, and stability of the delivery system. Team development, combining subject-matter experts (SMEs), graphics artists, learning specialists, designers, programmers, and managers, should make the development process more efficient and productive. This view assumes awareness of another lesson taught by the pioneers: that technology per se does not automatically change the trainees' experiences, but careful instructional design, quality materials, flexible delivery, timely tutor help, and convenient communications systems may.

Because of these differences, trainers in Web-based environments may find their duties and priorities changed. A study (National Education Association, 2000) of U.S. college instructors in a variety of institutions found the following:

- Most reported Web-based teaching was more personally rewarding than traditional methods: distance methods were seen as giving trainees better access to information, better quality materials, more help in mastering the subject matter, and more allowance for individual needs.
- Most instructors had at least some one-on-one contact with learners, and those who had reported higher levels of satisfaction in their teaching.
- Teaching at a distance required more instructor time; unfortunately, most training organizations did not formally recognize this fact.

- Despite lack of organizational recognition of the greater time required, almost three-quarters of the survey respondents held positive feelings about distance teaching (only 14% reported negative feelings).

The WBT Environment

As noted, if designed to do so, and if the trainer supports the shifts implied, WBT can change the basic relationship between the trainee and the tutor, and can alter fundamental characteristics of the training environment: outside information resources and a wider range of human contacts can be accessed; there are more choices and options for trainees (but trainees need greater maturity to exercise them wisely); and WBT environments can emphasize collaboration over competition (Relan & Gillani, 1997). Materials design, instructional methods, and "best" teaching practices are also affected in WBT environments.

Materials and Instructional Activities

In traditional face-to-face teaching, instructional materials may be prepared at the last minute, or even simply dispensed with (trainees being required to take notes from the comments and often random chalkboard musings of the instructor). In WBT, materials preparation is a major stage in program development. Complete WBT materials are self-contained, including organizers and instructions, with guidance and feedback provided through embedded questions and other self-evaluation activities. Support and orientation are available for any technologies used.

In well-designed and -managed WBT, instructional activities and materials may employ some or all of the following principles:

Typically, a wider range of resources, some from outside the local environment, is used.

Training may incorporate experiential learning and simulations, accessed via Web links.

Collaboration replaces competition.

The instructor is more a guide and coach than a dispenser of information.

Problem- or case-based learning is more common, sometimes increasing the training's realism and authenticity. (Students are permitted, even required, to clarify and refine questions themselves, without constant reliance on the trainer.)

Personal knowledge and experience are valued and included in problem-solving activities (Newby, Stepich, Lehman, & Russell, 2000).

"Best Practices" in WBT

Instructors in well-planned WBT adopt specific training strategies known to enhance learning. One training model recommends trainers strive for a balance between *interpersonal rapport* and *intellectual excitement*, requiring the trainer to be interpersonally warm, open, predictable, and learner-centered, while also being clear and enthusiastic about the training content (Lowman, 1994).

Another well-respected model of best practices recommends these training behaviors (Chickering & Gamson, 1989):

Encourage contacts between trainees and instructors.

Develop reciprocity and cooperation among trainees.

Use active learning techniques.

Give proper and timely feedback.

Emphasize time-on-task.

Communicate high expectations.

Respect diverse talents and ways of learning.

Bloom's (1984) classic description of the "alterable variables" of learning also provides guidance for Web-based trainers. Research in mastery learning showed that the following variables, when emphasized, produced learning outcomes similar to what could be achieved under ideal training conditions (one-to-one tutorial):

Provide well-designed tutorial instruction.

Give timely reinforcement.

Give appropriate and sensitive corrective feedback.

Provide cues and explanations, as needed.

Encourage learner classroom participation.

Assure trainees make effective use of time on task.

Help trainees improve reading and study skills, as required.

Individual Differences

One of the major differences between Web-based and more traditional forms of training is WBT's capacity for accommodating the individual expectations and preferences of trainees. This feature can be particularly valuable in meeting "special" needs, or those based upon adult trainees' *personal* and *situational* variables. Personal variables include age, maturity, personal health, time availability (and management skills), motivation, previous learning, financial circumstances, and life and developmental stages. Situational variables include factors such as location (related to the location of any required site-based training), admission and training program requirements, availability of counseling and advisement services, and personal issues such as transportation, health, and child-care (Cross, 1981).

WBT's capacity to accommodate differences effectively partially depends upon the trainees' capacity and willingness to exercise independence, autonomy, and self-direction. Even if trainees are adults or mature adolescents, the presence of the needed skills and maturity for self-directed learning cannot always be assumed. Trainees must be willing to exercise self-direction and independence in learning.

Problems arise in WBT situations when there is a mismatch between the self-direction the learning system permits and the expectations of the trainees. Mismatches between *teaching or training style* and *learning style* can result in dissatisfaction with the learning experience, or worse (dropout, failure). Programs are more successful if aligned with the developmental stages of individual learners. Trainee readiness may range from nearly

complete dependency to fully autonomous self-direction, requiring the trainer to function variously as an authority or coach, a motivator or guide, a facilitator or mentor, and, at the highest levels of self-direction, a consultant (Grow, 1991). Failure to provide learning conditions that align with trainees' expectations for support, interaction, or recognition may be one of the principal reasons for unacceptably high dropout rates in some WBT programs, a problem which, though hard to describe on a national basis, has been identified as a serious one for some online training (Frankola, 2001).

Economic Factors

The economics of WBT, though changing rapidly in the details, continue to directly impact training providers and consumers.

For Providers

Costs of development of WBT vary dramatically. Text-based WBT, involving conversion of existing material and using one of the many authoring tools available, may be economically accomplished by a subject-matter expert (SME) with basic instructional design skills. On the other hand, development of an hour of computer-assisted learning (CAL) using high-level authoring languages might require 40 to 150 hours (Szabo, 1998), and one complex 4- to 6-hour multiunit module in weather forecasting, incorporating multimedia and simulation effects, reportedly required a year, involved a team of instructional design and subject matter specialists, and ultimately cost $250,000 (Johnson, 2000). Financial considerations are primary in most WBT implementations: if an organization cannot afford the attendant costs (especially the often heavy initial investment in development), it may not be able to make the transition to WBT, even if the need is clear and the organization willing.

The financial case for WBT depends largely upon the relation of *fixed* to *variable* costs of development and delivery. Fixed costs are those incurred whether the training materials serve a handful or thousands of users. Fixed costs include staff salaries, equipment, and other capital costs directly related to development, including rent and other overhead costs. Variable costs are those that increase in relation to demand, such as printing, materials reproduction, and shipping; wages of section lecturers or lab demonstrators hired in response to registrations; additional clerical assistance; costs for licenses or copyright based on usage; and equipment for training delivery, which might have to be acquired to serve increasing demand. In WBT development projects, financial viability often depends upon fixed costs being kept to a minimum, and as many costs as possible remaining variable (dependent upon, and thus paid for, by demand).

Providers of WBT must promote their programs and services without overselling them. While WBT offers the *potential* for substantial convenience increases and improvements in efficiency (including reduced training costs), it is important to acknowledge that WBT results cannot be guaranteed to be uniformly or automatically better for all users. This is due to interactions among economic, technical, and organizational factors, and because of the importance of the design to the quality of the implementation. In fact, one of the paradoxes of the past decade's use of technology generally, including training applications, has been the persistent finding of "no significant difference" in training results, and the "productivity paradox," the failure of some industries to achieve economic benefits from technology implementations, while others made impressive gains, mainly through enhanced performance compared to the competition (Fahy, 1998). Nevertheless, where design converges advantageously with needs, opportunities, and a willing corporate culture, WBT has been proven successful (Vaas, 2001; Welsch, 2002).

For Consumers (Trainees)

WBT trainees have come to expect that they will have access to timely, economical, high-quality, self-paced training, virtually anytime and anywhere (Vaas, 2001). The keys to meeting these expectations are the *cost* and *accessibility* of the WBT technologies used (Bates, 2000), and relevance of WBT's interaction capabilities to specific user needs (Fischer, 1997).

For trainees with "special" learning needs, WBT's interaction capabilities can provide important advantages. The mobility-handicapped, those with learning disabilities (LDs), or attention-deficit disorders (ADDs), including ADDults (Keller, 1999), often find an environment with more learner controls, such as is typically found in WBT, helpful. Two core features of WBT directly applicable to special needs trainees include

- Structure—Advance organizers; clearly stated objectives, schedules, and timelines; embedded comprehension checks; integrated media under learner control; multimodal presentations; user-accessible performance records and reports; and communication links with the trainer and other support resources.
- Flexibility—Any place, any pace access.

For the physically handicapped or mobility impaired, the following features and characteristics of WBT can be of value:

- Distributed—Available in accessible locations.
- Interruptible—Trainees can take breaks when needed.
- Modular—Single or multiple skills can be addressed.
- Multisensory—Sight, sound, and tactile cues can be incorporated.
- Nonlinear—Presentation sequence can be varied.
- Portable—Easily moved or transferred.
- Responsive—Adoption time is relatively short.
- Transferable—Crosses cultural, language, situational, physical, and geographic barriers (Gerofsky, 1998).

Those working with special needs audiences have the capability to monitor progress regularly (including administering testing as needed), communicating with trainees easily, allowing trainees to communicate with each other (socialize), and helping coordinate and support training components (including the efforts of helpers and support staff).

Trainees without special needs can benefit from WBT, too; many of the above features can assist most trainees

to complete their training with more satisfaction or less stress, by reducing potential conflicts with trainees' careers and personal lives. Experience has shown that training may precipitate problems among trainees in general, which skillful use of WBT's own communications capabilities may help solve, such as

- Feelings of inadequacy at the sheer amount of material to be covered;
- Delays in receiving feedback or answers to questions;
- Keeping up with the variety of discussions and interaction often present in online (computer-mediated communications, CMC) discussions;
- Adjusting to the absence of visual clues in group relations; and
- Fatigue and health problems arising from reliance on unfamiliar technologies (eye-strain or posture problems at the computer, for example).

THE FUTURE OF WBT

While WBT has proven its potential value for training delivery, there are barriers that may restrict or slow its expansion, chiefly bandwidth availability, security and privacy issues, and user access to required technologies.

Bandwidth and Security

Present "POTS" Systems

Today, most home Web users still access the Internet using POTS (plain old telephone system) dial-up modems. While common, cheap, and reliable, the transfer speeds available with this technology are a limiting factor in the evolution of multimedia-based WBT.

Typical POTS Internet connections move data (theoretically; actual rates are always lower) at 56 Kbps (kilobits per second); at the higher end of the bandwidth spectrum are speeds of 1,000 to 1,500 Kbps (cable, DSL, ISDN, T-1), and some that exceed 2,500 Kbps (RADSL, T-3). To compare these speeds with actual requirements of a common medium, TV graphics (depending upon the screen resolution of the receiver) require throughput of from 1,800 to over 110,000 kilobits of data *per second*. None of the current POTS- or cable-based transmission methods are capable of more than a small percentage of that rate.

Considerably less than full TV-quality video could be adequate for many WBT applications. However, training organizations appear to be ignoring this fact: a survey of the intentions of U.S. colleges and universities showed that, between 1995 and 1998, 2-way video use grew 22%, while many of those institutions that had not yet committed to it were planning to do so within 3 years. On the other hand, use of audio-only delivery (in the forms of one-way audiotapes, or two-way teleconferencing or VOIP [Voice over Internet Protocol; Internet audio] connections), potentially very powerful and relatively low-cost media, remained virtually unchanged at about 10% of institutions (U.S. Department of Education, 1999). A similar focus on high-bandwidth applications was found in the private sector: a 2001 survey showed that "live video" was expected to grow from 7% to 31% in the next year ("Real-time help," 2001).

Satellite Systems

Satellites provide a powerful alternative to ground-based data transmission in WBT systems: satellite-based signals can be broadcast over a much wider area than broadcast tower-based or independent wire/cable transmission. All the usual production costs apply to satellite-based delivery; in fact, production costs may be higher, since the greater potential audience may warrant the highest production values. In addition to the costs of developing and launching the satellite the potential for equipment failure is significant, given the inconvenience of service calls.

At present, broadcast delivery systems, including satellite, are federally regulated in North America, but if deregulation of these services occurs, cable TV and telephone systems will be able to compete to deliver each other's services (as they already do in other parts of the world). The impetus for deregulation is the fact that video and audio signals, when digitized, are simply data, which can in principle (if not yet in law) be delivered by anyone with data transmission capabilities.

Wireless Systems

Wireless technologies (other than satellites) are also changing rapidly, with direct implications for training. The cost of installing "fixed" wireless capabilities (the short-range 802.11 protocols) in existing buildings has fallen below the cost of wiring (or rewiring). Besides cost savings, wireless technologies are quick to install, and are highly portable, so they can be readily moved as opportunities or demands change. Portability also permits cultivation of users by making equipment and services quickly available to those best prepared to make good use of them (McKenzie, 1999). In late 2001, 7% of colleges and universities in the United States had campus-wide wireless installations, and 51% of institutions reported some wireless capabilities, up from 30% in 2000 (Campus Computing Project, 2001). The cost and availability of other types of wireless technologies with much wider coverage areas, such as cell phones and IM (instant messaging) devices, are also dropping, though the training capabilities of these technologies have not yet been widely tested.

Despite cost reductions, wireless systems in training have some major disadvantages: transmission speeds are typically slower; interference may lead to transmission errors, further reducing speed, especially if other electronics operate in the same environment; range depends upon the site layout and configuration of the network; and (as discussed later) wireless systems are much more vulnerable to security breaches.

Web Availability

Another area of concern for the future growth of WBT is trainee access to the Internet. U.S. Census data provide a view of Internet access nationally (U.S. Census Bureau, 2001):

Access, though increasing, is not yet universal: in the United States by 2000, more than half (51%) of households had access to a computers (up from 42% in December 1998), but Internet access was lower, and differed by regions: highest levels were found in the West, where 40.7% of households had Internet access,

while the lowest were in the South, where the level was 34.3%.

Ethnicity was related to Web access: white households were almost twice as likely to be connected as nonwhite households.

Income was also associated with Web access: families with annual incomes over $75,000 were three times more likely to be online than those with incomes below $24,000.

Urban access rates were more than 12% higher than nonmetropolitan rates.

The presence of a child in the family increased the likelihood of both computer ownership (by almost 22%) and Internet access (by over 16%).

Age was a major factor: the group most likely to own a computer and be connected to the Internet was aged 25 to 44 (ownership, 61.0%; Internet, 50.2%), followed by the cohort aged 45 to 64 (56.9% and 46.7%, respectively); least likely was the 65 and older group (at 24.3% and 17.7%). In Canada, similar patterns were found ("Getting connected," 1999).

Gender has historically been an issue in Web access, though experience with WBT appears to considerably reduce, and may even be reversing, former access patterns (Wark, e-mail, September 26, 1999). Traditionally, men have been more engaged generally with all aspects of computer use than women; women have reported finding CMC less personal and online environments less comfortable than face-to-face interaction, and have consequently been more reluctant to enroll in computer courses (Blocher, 1997; Kirkpatrick & Cuban, 1998).

The Internet may be about to change that pattern: since May 2000, trends have shown that women as a group exceeded the number of men online, so that by June 2001 women comprised 40.9 of Internet users and men 39.8%. Interestingly, in relation to the question of the feasibility of the Internet for the training of older workers and women, the largest increase in usage was among individuals over age 35, including women in that age group. Even women who were mothers increased their Internet use: mothers who were online averaged 16 hours 52 minutes per week, more even than online teens (who averaged just over 12 hours weekly). Another fact with training implications: mothers who might be expected to have the least free time (single mothers and those with 3 or more children) were online the most, averaging about 20 hours per week, 20% more than the overall average (Saunders, 2002). While women as a group were increasing usage significantly, men continued to lead women in frequency and intensity of Internet use: when online, men averaged 16% more time online than women, viewed 31% more pages, and logged-on 11% more often (Pastore, 2001).

Assuring Security and Privacy

WBT systems may be victimized from the outside by viruses, unauthorized intrusions, sabotage, or fraud. It is sobering that 75% of security breaches in the private sector (incidents of sabotage, hacking, or data theft) are committed by the institution's own personnel (unauthorized present employees or former staff). Wireless operations are particularly vulnerable: not only are wireless systems technically more difficult to secure, wireless users tend to be more casual about employing available security measures (Miller, 2001).

Computer viruses of all kinds have become almost ubiquitous and are expected to grow in number. By one estimate, in 2001 viruses were found in 1 of every 300 e-mail messages; at the current rate of proliferation, by 2008 the ratio will be 1 in 10, and by 2013 it will be 1 in 2 ("Outbreak," 2001). "Virus" is the generic name for all malicious programs, including worms and Trojan horses; the term *malware* has been suggested for all these malicious forms of code (Seltzer, 2002). Worms and Trojans are special forms of malware, in that they programmed to spread by themselves without human intervention (usually by e-mail), while simple viruses require individuals to deliberately share files for the viruses to be able to move from machine to machine. The magnitude of the threat from malware has led most organizations to install protection in the form of virus detectors and firewalls. As well as preventing unauthorized intrusions from the outside, the latter restrict the access of those behind the firewall to outside people and materials (for example, VOIP may be impossible, or severely limited). While protecting the security of those behind them, firewalls can constitute a major barrier to WBT generally.

Confirming the identity of trainees is also a potential problem in WBT, especially if trainees do not routinely meet face-to-face with instructors. Just as failure of system security might expose a training organization to embarrassment, expensive down-time, or even litigation, failure of an organization's screening and monitoring systems leading to a fraudulent registration, award of credit, or granting of a credential might be disastrous for its reputation. Fortunately, technologies exist and are becoming more economical to help with trainee identification and authentication: biometric devices such as voiceprint and fingerprint identifiers, and remote cameras, have recently become available for economically checking identities of trainees and supervising remote testing events (Miller, 2001).

Implementing WBT

In order to produce high-quality WBT results, incorporating flexibility, efficiency, and individualization, training organizations must assure that certain elements are present. Among these are a conducive social learning environment and institutional collaboration to assure efficient provision of courses and transfer of credit, including prior learning assessment.

Cohort Learning and Socialization

All group learning is social. WBT technologies provide the option for interaction and collaboration, which should in turn increase "social presence" (Garrison, 2000) and reduce "transactional distance" among participants (Moore, 1991).

Cohort-based WBT provides a supportive social network, especially for adult trainees who may lack confidence or who may face hurdles in returning to formal learning. Cohort-based programming uses more active,

cooperative, and collaborative learning strategies than more traditional methods. The cohort structure consists of a group of trainees who enter and complete the program together in a predetermined and prescheduled series of common training experiences. Trainees may meet face-to-face occasionally, even in primarily distance programs, to initiate social interaction that technology-based interaction can then sustain. The resulting training, though lockstep, appears to be successful in creating trust, empowerment, and support, while reducing adjustment problems and drop-out, especially among older trainees (Saltiel & Russo, 2001, p. vii).

The communications capabilities of WBT technologies can also reduce isolation, and increase motivation and social interaction. Computer-mediated communications includes one-on-one interaction (e-mail), one-to-many connections (conferencing, list-servs), data sharing (file attachments), and information access (via the Web's links). Computer conferencing is a potentially powerful means of creating community in WBT programs. CMC may increase comprehension of training objectives by promoting peer-to-peer interaction, but basic ground-rules enforced by a conscientious moderator are required to assure that the resulting CMC interaction is effective, and to help avoid asocial outcomes such as "social loafing." (Other forms of asocial interaction, such as rudeness or "flaming," are rare in moderated training interactions, but may occur in public nonmoderated environments such as list-servs.)

Institutional Collaboration

WBT assists trainers to address the globally recognized need for more efficient and flexible training delivery, including transnational standards, increased quality assurance (based on competency-based curricula), multiskill training, and the appropriate adoption of electronic technologies to increase trainee success. Driving this is the fact that employers have historically not been very satisfied with what they perceive as the public school system's inflexibility, and apparent inability to prepare its young graduates better for employment.

Employers expect graduates to be capable of teamwork, creativity, problem solving, and adaptability. To counter the lack of flexibility in institutional training, some employers advocate training-on-the-job (TOJ) programs. TOJ is seen as providing a better training experience overall, especially if linked with WBT opportunities. The combination has been regarded as better addressing employees' convenience and privacy, while allowing employers to monitor relevance, and permitting the involvement of experienced employees in the training of novices (Conference Board of Canada, 2001).

Prior learning assessment (PLA), like transfer credit, benefits trainees who have accumulated credits over time, perhaps from a variety of sources, without ever completing a credential. PLA recognizes that learning may appear to be haphazard, while yet equipping the trainee with skills and knowledge worthy of formal recognition (especially if accompanied by relevant work experience). WBT may help a trainee integrate and complete a program based on PLA credits, a process sometimes called "cap-stoning."

	SAME TIME	DIFFERENT TIME
SAME PLACE	1	2
DIFFERENT PLACE	3	4

Figure 1: Diagram attributed to Coldeway by Simonson, Smaldino, Albright, & Zvacek (2000, p. 7). A similar typology is also found in Johansen, Martin, Mittman, Saffo, Sibbet, & Benson (1991, p. 17.).

Media and the Future

Some technologies require broadband (simulations, full-motion video, high-quality audio); on the other hand, some technologies are useful only for relatively limited training purposes (text on paper is ideal for information transmittal, but lectures and collaborative group sessions are poor vehicles for this purpose). Research has demonstrated that the impact of technologies on training outcomes depends upon specific media characteristics, and that technologies differ in respect to their cost, accessibility, teaching implications and impacts, interactivity, user-friendliness and control, organizational impact, novelty to users, and speed of adoption and adaptation (Bates, 2000).

Knowing this, WBT programmers can make better decisions about the "right" technology for a particular application on the basis of the amount of separation (the "distances") between trainees and elements of their WBT programs. Figure 1 illustrates how time and place of training can vary in WBT.

Differences in the time and place of training can impact trainees, and affect the flexibility of the training, in several ways:

Quadrant 1 ("same time, same place"). Training in this quadrant is *synchronous* (same time) and *site-bound* (a classroom or lab is set aside for it). Participants in the same place at the same time may still experience "distances"—psychological, interpersonal, sociocultural, linguistic, philosophical, etc.—which may create barriers to communication and learning, requiring timely trainer intervention.

Quadrant 2 ("same place, different time"). Training is site-based, but permits asynchronous access in the form of correspondence modules or packages at a training or learning center. Regular attendance at the designated training site is often required so progress can be monitored.

Quadrant 3 ("same time, different place"). Training is synchronous but not site-bound; technology permits access from off-site, but only at "class" time. Trainees need appropriate remote-access media.

Quadrant 4 ("different time, different place"). This is "any pace, any place" training. Requires the training institution to provide materials, support, interaction opportunities, and administrative arrangements at the trainees' convenience. Trainees require appropriate remote-access technologies.

The Evolution of Present Media

Training media are changing as bandwidth improves, and as they do, new forms of familiar technologies are presenting WBT designers with options and capabilities previously unavailable or not cost-effective. Examples include the previously mentioned VOIP, and reusable learning objects, with (for those capable of accessing it) multimedia.

VOIP technology permits use of computers for voice communications, either as person-to-person private conversations or in multipoint group sessions. Users commonly require no more than normal dial-up access to the Web, plus a sound card, microphone, and speakers. For video, VOIP typically provides a $2'' \times 3''$ display, with a refresh rate dependent upon bandwidth. (IP video is presently not of full-motion quality; its jerkiness and grainy nature lead some users to switch it off completely, relying upon audio alone.) The quality of IP audio is usually adequate to good. VOIP services are very cheap or even free, and the quality and reliability are improving rapidly.

Standards and Reusable Learning Objects (RLOs)

Training materials are increasingly designed to be reused ("repurposed"), as standards are developed by organizations such as the Aviation Industry CBT Committee (AICC), the Instructional Management Systems (IMS) Global Learning Consortium, the Institute of Electrical and Electronics Engineers (IEEE), and the Advanced Distributed Learning (ADL) Initiative of the Department of Defense (developer of the Shareable Courseware Object Reference Model [SCORM]) (Hodgins & Conner, 2000). WBT instructional materials developed under RLO standards are portable, for maximum present and future (re)use. *Metadata tagging* of RLOs attempts to assure several outcomes relevant to potential WBT uses:

- Flexibility—Material is designed to be used in multiple contexts and to be easily reused in other applications.
- Ease of update, search, and content management—Metadata tags allow quick updating, searching, and management of content through sophisticated filtering and selecting capabilities.
- Customization—Modular learning objects enable more rapid program development or revision.
- Interoperability—RLOs are designed to operate on a wide range of training hardware and operating systems.
- Competency-based training—Competency-based approaches to training are promoted by RLOs; materials are tagged to competencies, rather than subjects, disciplines, or grade levels.

Developing training materials to RLO standards increases their value by making reuse easier and thus more likely. Also, the presence of standards means savings may be realized in reduced design and development time, and in some cases revenue may be generated from sales (Longmire, 2000). Multimedia RLOs must be used sparingly, if at all, in training intended for home use, however: while in late 2001 92% of U.S. businesses with 1,000 employees or more reportedly had broadband connections (T1, T3, DSL, cable, ATM, frame-relay or faster), fewer than 20% of home users had these capacities (Metz, 2001).

The "New" Internets

Worldwide growth in commercial use of the public Web has resulted in ongoing efforts to replace it with a version reserved for academic and military use. In North America, new Internet developments include the United States' Internet2, the Information Technology Research and Development (ITR&D) Program, and Canada's CA*net system.

Internet2 is a collaborative effort of more than 120 U.S. universities, partnering with industry and government (through the National Science Foundation) to create an environment that "is not clogged with music, commercial entities and porn" (Rupley, 2002). As with Canada's CA*net II, CA*net 3, and CA*net 4 (launched in mid-2002), research and education are the focus (Networking, 2002b). In mid-2001, the first implementations of Internet2 in U.S. K–12 schools had commenced (Branigan, 2001), with public access available in early 2002.

The Information Technology R&D Program continues and broadens the agenda of the NGI (Next Generation Internet; NGI, 2001) project, a multiagency, federally sponsored research and development program that, by the time it concluded in 2001, had helped achieve major advances in networking speeds. The ITR&D program includes the collaborations of 12 very high profile agencies (for example, the National Science Foundation [NSF], NASA, the Defense Advanced Research Projects Agency [DARPA], the National Institutes of Health [NIH], and the Environmental Protection Agency [EPA], among others). The ITR&D program conducts research and development in various "program component areas" (PCAs), including cutting-edge projects in high-end computing, scalable information infrastructure, large-scale networking R&D, and high-confidence software and systems. The commitment is significant: the budget requested for 2002 was almost $2 billion (Furlani, 2002).

The California Institute of Telecommunications and Information Technology (http://www.calit2.net) consortium plans to build a high-speed *wireless* network that is cheap, always on, and accessible through a variety of technologies. "Tether-free" online technologies are currently represented by palm-size computing and cell phone-type communications devices. The evolutionary potential for training applications of these highly portable technologies is seen as tremendous, assuming concurrent advances in AI (artificial intelligence) systems and voice command capabilities. Public funding for this project reached $300 million in mid-2001 (Chapman, 2001).

CA*net II, Canada's first "next generation" Internet initiative, demonstrated that a dedicated, noncommercial

research Internet was feasible and needed by Canadian academics. Although initially established as a public enterprise CA*net II was privatized in 1993, then coming under the direction and control of CANARIE (the Canadian Network for the Advancement of Research, Industry and Education), a consortium of private Canadian organizations and academic institutions.

The purpose of the CA*net program was to upgrade Canada's research and development infrastructure, especially networking and communications capabilities, and to permit joint ventures and collaborations to promote the involvement of Canadian business and industry in the knowledge-based economy. Like the United States' NGI and Internet2 systems, CA*net uses "special access points" in each Canadian province to provide access to the system's high-capacity, high-speed communications link called the vBNS (very high-speed Broadband Network Service), which in turn connects to the network's backbone.

Canada's CA*net II system, the world's first nation-wide network of its sort, was succeeded by CA*net 3 (*"Hooked to CA*net 3,"* 2000), and, in 2002, by plans for CA*net 4. Like all high-capacity networking systems, CA*net uses different wavelengths (colors) of light and fiber optics to permit dozens of different transmission "channels" (CA*net II had only one). CA*net 3 was some 20 times faster than CA*net II, and 50,000 times faster than present commercial Internet services; CA*net 4 is expected to continue this evolution in high-speed networking, improving the potential for quality WBT (Networking, 2002a, 2002b).

CONCLUSION

Predicting the future for WBT is, on one hand, simple: it will expand, grow, and become central to more forms of future training worldwide. That prediction is safe, both because so much has already been invested in the Internet delivery infrastructure and because the Web has already proven to be so effective in reducing direct training costs, increasing access, and addressing serious skills deficiencies in an increasingly competitive and technological world.

On the other hand, no one can predict, even a short time into the future, specific details of the training advances WBT will help make possible in the future. Greater usage does not imply standard use; more investment may not result in quality increases; reliance on WBT does not guarantee enthusiasm or success by specific groups of trainees. Experience with other forms of training delivery have demonstrated that instructional design, perhaps even more than the medium, makes training effective, and trainee support and overall relevance of the content make it satisfying to users.

While WBT cannot guarantee that future training will not be pedestrian or inefficient, its potential strengths may suggest how training of all kinds might be improved. The core elements of the Web as a training device—ubiquity, accessibility, stability, supportiveness, redundancy, and friendliness—are common to other successful teaching environments. For WBT to expand, developers, trainers, trainees, and employers must see these aspects of the Web as potential assets for their training programs. If the inherent features of the Web are seen as important to training, continued growth and expansion of WBT are certain to occur.

GLOSSARY

Asynchronous "Different time" (and often different place, i.e., the trainees' workplace or home) training and communications interactions, made possible by technologies that collect messages and make them available at the convenience of the reader. Examples include e-mail, CMC, and list-servs.

Attention deficit disorder (ADD) A form of learning disability that, in children, results in short attention span and lack of focus, sometimes accompanied by hyperactivity; ADDults, adults with ADD.

Bandwidth The capacity of a channel (for example, a telephone line or a coaxial cable) to carry data. High bandwidth permits multimedia (audio, video, animation), while low bandwidth limits users to text or simple graphics, and may preclude VOIP and other new interaction tools.

Computer-mediated communications (CMC) Text-based, asynchronous communications, usually restricted by password to a designated group such as a class or training cohort, and moderated to assure that the discussion stays on topic and is civil.

Internet The public network of linked computers accessible by anyone with a Web browser.

Intranets Private computer networks that may or may not also provide Web access, providing security and control not available on the public Internet by limiting access and controlling the content available to users.

Malware Any form of malicious code intended to infect a computer or a network, including viruses, worms, and Trojan horse programs (Seltzer, 2002).

Metadata The identifying material added to RLOs to permit easy reuse or "repurposing."

Online Training that includes potential for synchronous or asynchronous electronic (often Internet-based) interaction between the trainee and the trainer, the training materials, and other trainees.

Prior learning assessment (PLA) A process intended to result in the granting of credit toward a credential for a trainee's accumulated formal and nonformal learning experiences; recognizes and rewards the learning of individuals who have not been able to complete a credential at one institution in the normal way, due to career or personal reasons.

(Reusable) learning objects (RLOs) Materials modularized, packaged, and labeled to encourage cataloguing, access, and reuse in multiple contexts.

Synchronous "Same time" and (often) same place training or communication interactions, e.g., face-to-face training.

Training Instruction in psychomotor skills and knowledge primarily for practical purposes and relatively immediate application.

Voice over Internet protocol (VOIP) The capability of Internet-based programs to provide voice point-to-point or point-to-multipoint connections to anyone with a browser, a sound card, microphone, and

speakers; may also include limited video or graphics capabilities

Web A synonym for the Internet (see above).

Web-based training (WBT) Includes training-on-the-job (TOJ) and workplace training, formal on-campus technical training, and elements of training or professional development (PD) conducted primarily face-to-face, but including some online components, with the Internet or an intranet as the access/delivery vehicle.

CROSS REFERENCES

See *Distance Learning (Virtual Learning); Internet Literacy; Voice over Internet Protocol (IP)*.

REFERENCES

Bates, A. W. (2000). *Managing Technological Change*. San Francisco: Jossey-Bass.

Blocher, M. (1997). *Self-regulation of strategies and motivation to enhance interaction and social presence in computer-mediated communication*. Unpublished Ph.D. dissertation, Arizona State University.

Bloom, B. S. (1984, June–July). The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher*, 4–16.

Branigan, C. (2001, April 16). Five states are first to offer Internet2 to schools. *eSchool News Online*. Retrieved June 2001 from http://www.eschoolnews.com/showstory.cfm?ArticleID=2535

Briggs, L., & Wager, W. (1981). *Handbook of procedures for the design of instruction* (2nd ed.). Englewood Cliffs, NJ: Educational Technology.

Burge, E. J., & Roberts, J. M. (1993). *Classrooms with a difference: A practical guide to the use of conferencing technologies*. Toronto: University of Toronto (OISE), Distance Learning Office.

Campus Computing Project (2001). *The 2001 National Survey of Information Technology in US Higher Education*. Retrieved May 2002 from http://www.campuscomputing.net/summaries/2001

Chapman, G. (2001). Consortium sets sights on a "new Internet." *Digital Nation* [online article and column, Los Angeles Times, April 5]. Available by listserv from chapman@lists.cc.utexas.edu

Chickering, A., & Gamson, Z. (1989, March). Seven principles for good practice in undergraduate education. *AAHE Bulletin*, 3–7.

Conference Board of Canada. (2001). *Performance and Potential 2001–02*. Retrieved May 2002 from http://www.conferenceboard.ca/pandp/documents/pandp.01.pdf

Cross, P. (1981). *Adults as learners*. San Francisco: Jossey-Bass.

Dewey, J. (1938). *Experience and education*. New York: Macmillan.

Dick, W., & Carey, L. (1978). *The systematic design of instruction*. Dallas: Scott, Foresman & Co.

Fahy, P. J. (1998). Reflections on the *productivity paradox* and distance education technology. *Journal of Distance Education, 13*(2), 66–73.

Fischer, B. (1997, January/February). Instructor-led vs. interactive: Not an either/or proposition. *Corporate University Review*, 29–30.

Fox, E. A. (2002). Multimedia, hypertext and information access. Course notes, CS 4624, Virginia Polytechnic Institute and State University. Retrieved May 2002 from http://ei.cs.vt.edu/~mm/

Frankola, K. (2001). Why online learners drop out. *Workforce*. Retrieved July 2002 from http://www.workforce.com/archive/feature/22/26/22/index.php

Furlani, C. M. (2002). National IT R&D Program. Retrieved May 2002 from http://www.itrd.gov/about/presentations_nco/2002/furlani_nhpcc_20020403/index.php

Gagne, R., & Briggs, L. (1979). *Principles of instructional design* (2nd ed.). New York: Holt, Reinhart and Winston.

Garrison, R. (2000). Theoretical challenges for distance education in the twenty-first century: a shift from structural to transactional issues. *International Review of Research and Open and Distance Learning (1)*1. Retrieved September 2000 from http://www.icaap.org/iuicode?149.1.1.2

Gerofsky, G. (1998, May). An Athabasca challenge: Providing for disabled learners at a distance. Paper presented at the 14th Annual CADE Conference, Banff, Canada.

Getting connected, staying unplugged. (1999). *Innovation Analysis Bulletin, 1,* 4. Statistics Canada–Catalogue No. 88-003-XIE. Retrieved March 2000 from http://www.statcan.ca:80/english/freepub/88-003-XIE/free.htm

Glaser, R. (Ed.). (1978). *Advances in instructional psychology*. Hillsdale, NJ: Lawrence Erlbaum.

Grow, G. (1991). Teaching learners to be self-directed. *Adult Education Quarterly, 41*, 125–149.

Harapniuk, D., Montgomerie, T. C., & Torgerson, C. (1998). Costs of developing and delivering a Web-based instruction course. In *Proceedings of WebNet 98—World Conference of the WWW, Internet, and Intranet*. Charlottesville, VA: Association for the Advancement of Computing in Education. Retrieved July 2002 from http://www.quasar.ualberta.ca/IT/research/nethowto/costs/costs.html

Hauben, M. (1994). *History of ARPANET*. Retrieved April 2002 from http://www.dei.isep.ipp.pt/docs/arpa.html

Hodgins, W., & Conner, M. (2000, Fall). Everything you always wanted to know about learning standards but were afraid to ask. *LiNE Zine*. Retrieved June 2001 from http://www.linezine.com/2.1/features/wheyewtkls.htm

Hooked to CA*net 3 (2000, March). *University Affairs*, 27.

Johansen, R., Martin, A., Mittman, R., Saffo, P., Sibbet, D., & Benson, S. (1991). *Leading business teams*. Don Mills, Ontario, Canada: Addison-Wesley.

Johnson, V. (2000, June). Using technology to train weather forecasters. *T.H.E. Journal Online*. Retrieved March 2002 from http://www.thejournal.com/magazine/vault/articleprintversion.cfm?aid=2880

Jones, M., & Farquhar, J. (1997). User interface design for web-based instruction. In B. Khan (Ed.), *Web-based instruction* (pp. 239–244). Englewood Cliffs, NJ: Educational Technology.

Kaye, A. (1989). Computer mediated communication and distance education. In R. Mason & A. Kaye (Eds.), *Mindweave* (pp. 3–21). Toronto: Pergamon Press.

Keller, F. S. (1968). Goodbye, teacher! *Journal of Applied Behavioral Analysis, 1*(1), 79–89.

Keller, V. (1999). *Adaptation and application of a transcript analysis tool to analyze a computer-mediated communication (CMC) distance education course transcript*. Unpublished master's thesis, Athabasca University, Athabasca, Alberta, Canada.

Kirkpatrick, H., & Cuban, L. (1998, July–August). Should we be worried? What the research says about gender differences in access, use, attitudes, and achievement with computers. *Educational Technology*, 56–61.

Leiner, B., Cerf, V., Clark, D., Kahn, R., Kleinrock, L., Lynch, D. C., Postel, J., Roberts, L. G., & Wolf, S. (2002). *A brief history of the Internet*. Retrieved May 2002 from http://www.isoc.org/internet/history/index.shtml

Longmire, W. (2000, March). A primer on learning objects. *Learning Circuits*. Retrieved June 2001 from http://www.learningcircuits.com/mar2000/primer.html

Lowman, J. (1994). What constitutes masterful teaching. In K. Feldman and M. Paulsen, *Teaching and learning in the college classroom* (pp. 213–225). Needham Heights, MA: Simon & Schuster.

Mager, R. (1975). *Preparing instructional objectives* (2nd ed.). Belmont, CA: Fearon.

McKenzie, J. (1999). Strategic deployment of hardware to maximize readiness, staff use and student achievement. *From Now On, 8*(8). Retrieved August 2000 from http://www.fno.org/may99/strategic.html

Metz, C. (2001). Broadband inside. *PC Magazine, 20*(20), iBiz1–iBiz8.

Miller, M. (2001). The broadband boom. *PC Magazine, 20*(17), 8.

Moore, M. G. (1991). Editorial: Distance education theory. *American Journal of Distance Education, 5*, 1–6.

National Education Association (2000). A survey of traditional and distance learning higher education members. Washington, DC. Retrieved August 2000 from http://www.nea.org/he/abouthe/distance.html

Networking (2002a). $100 million in federal budget for CA*net 4. *Networking, 6* (1). Retrieved May 2002 from http://www.thenode.org/networking/january2002/briefs.html#1

Networking (2002b). CA*net 4 coming down the pipe. *Networking, 6*(5). Retrieved July 2002 from http://www.thenode.org/networking/may2002/briefs.html#1

Newby, T., Stepich, D., Lehman, J., & Russell, J. (2000). *Instructional technology for teaching and learning* (2nd ed.). Upper Saddle River, NJ: Merrill.

NGI (2001). Next Generation Internet (NGI) Initiative. Retrieved March 2003 from http://www.ngi.gov/sc99/fast_facts.html

Outbreak (2001). *PC Magazine, 20*(21), 30.

Pastore, M. (2001). Women maintain lead in Internet use. Retrieved July 2002 from http://cyberatlas.internet.com/big_picture/demographics/article/0,,5901_786791,00.html

Piaget, J. (1952). *The origins of intelligence in children* (trans.). New York: International Universities Press.

Real-time help (2001) *PC Magazine, 20*(6), 73.

Relan, A., & Gillani, B. (1997). Web-based instruction and the traditional classroom: similarities and differences. In B. Khan (Ed.), *Web-based instruction* (pp. 41–46). Englewood Cliffs, NJ: Educational Technology.

Rupley, S. (2002). I2, near you. *PC Magazine, 21*(4), 21.

Rutherford, S., Patrick, J., Prindle, L, & Donaldson, S. (1997). Computer-managed learning (CML), an introduction. Edmonton, Canada: Grant MacEwan Community College and Northern Alberta Institute of Technology.

Saettler, P. (1990). *The evolution of American educational technology*. Englewood, CO: Libraries Unlimited.

Saltiel, I. M., & Russo, C. S. (2001). *Cohort programming and learning*. Malabar, FL: Krieger.

Saunders, C. (2002). Moms, Hispanics increasing web use. Retrieved July 2002 from http://cyberatlas.internet.com/big_picture/demographics/article/0,,5901_1041581,00.html

Seltzer, L. (2002). Personal antivirus. *PC Magazine, 21*(11), 96–103.

Simonson, M., Smaldino, S., Albright, M., & Zvacek, S. (2000). *Teaching and learning at a distance.* Upper Saddle River, NJ: Merrill.

Skinner, B. F. (1971). *Beyond freedom and dignity.* New York: Bantam Books.

Suppes, P. (1978). The role of global psychological models in instructional technology. In R. Glaser (Ed.), *Advances in instructional technology* (Vol. 1, pp. 229–259). Hillsdale, NJ: Lawrence Erlbaum.

Szabo, M. (1998). *Survey of educational technology research*. The Educational Technology Professional Development Project (ETPDP) Series. Edmonton, Alberta, Canada: Grant MacEwan Community College and Northern Alberta Institute of Technology.

Thorndike, R. L. (Ed.). (1971). *Educational measurement*. Washington, DC: American Council on Education.

U.S. Census Bureau (2001). *Home computers and Internet use in the United States*: August 2000. Retrieved July 2002 from http://www.census.gov/prod/2001pubs/p23–207.pdf

U.S. Department of Education, Office of Educational Research and Improvement (1999). *Distance education at postsecondary education institutions*. Retrieved April 2002 from http://nces.ed.gov/pubs2000/2000013.pdf

Vaas, L. (2001). The e-training of America. *PC Magazine, 20* (22), ibiz1–ibiz4.

Wagner, E. (1994). In support of a functional definition of interaction. *American Journal of Distance Education, 8,* 6–29.

Warren, A., Brunner, D., Maier, P., & Barnett, L. (1996). *Technology in teaching and learning.* London: Kogan Page.

Welsch, E. (2002). Cautious steps ahead. *Online Learning, 6,* 20–24.

Webcasting

Louisa Ha, *Bowling Green State University*

INTRODUCTION

Definition of Webcasting

Webcasting is the delivery of media contents and any digital information in various formats such as texts, graphics, and audio and video files on the World Wide Web to Internet users. This definition states the two characteristics of webcasting:

(1) The recipients of webcasts must have Internet access to receive the content.
(2) The content of webcasts can range from simple text to rich media files with multimedia capabilities.

Webcasting includes unicast (which serves a multimedia file in real time to a single user) and multicast (which allows many users to receive the Internet data streams at the same time with special software and hardware installed at different connections on the Internet).

There are many other terms commonly used to refer to webcasting as well. For example, cybercasting is synonymous with webcasting. Another common term is Internet radio or Web radio. However, Web radio refers to one type of webcasting that uses a radio station format and provides primarily audio files. Some of the Web radio stations only broadcast on the Internet, while others broadcast content from their offline radio station counterparts. Apart from audio content, Internet radio stations display texts and visuals such as photos of the show hosts and the playlist of the show. Several sites such as Yahoo!'s Launch (http://launch.yahoo.com/), flip2it.com (http://www.flip2it.com/schedule.asp?webcaster), Angelfire Radio (http://www.angelfireradio.com/search.html), and Classical Live Online Radio (http://www.classicalwebcast.com/start.htm) provide listings of Internet-only webcasts.

Reasons for Using Webcasting and Significance of Webcasting to the Internet and E-commerce World

Webcasting offers many benefits to individuals and organizations that need to disseminate information and content. The interactivity of computers allows the personalization and customization of information to consumers with great ease. Webcasting can deliver contents to the mass audience via the Internet and also to targeted audience via an intranet, an extranet, or on a subscription basis on the Internet. In addition, the Internet can transmit contents instantly to anywhere with Internet access, and if only nonmultimedia content is being webcast, the cost of webcasting is generally lower than the cost of a television or radio broadcast. Video webcasting, however, can be an expensive proposition because of the large server space and bandwidth required. The program length, audience number, and transmission speed determine how much server space and bandwidth are needed. The cost of delivering one 30-minute video webcast to 100,000 people can run from US$12,500 for a live webcast and US$9,500 monthly for an on-demand webcast. This price does not include the costs of overhead, production, promotion, and marketing involved to make a webcast successful.

Mass media, businesses, and individuals who want to have voices on certain issues and provide electronic media content to a geographically diverse audience use webcasting as an alternative content delivery medium. To established media such as broadcast and cable TV networks, webcasting means opening the battlefront to another new medium against competitors and increasing the value of their media content. Webcasting also means less reliance on intermediaries such as cable system operators or local broadcast TV stations (Ha and Chan-Olmsted, 2001). To

radio stations and broadcast TV stations, which are basically local operations, webcasting means the disappearance of geographic coverage barriers. To newspapers and other print media, their provision of webcasting transforms their status to a multimedia content provider.

To companies, webcasting opens new doors to how training is conducted and information is shared. Instead of restricting the training to a particular time at a particular location, employees can retrieve training materials such as demonstration videos and conference sessions at their own office or home at a convenient time. Current and prospective customers can retrieve product demonstration videos, virtual company tours, and other interactive multimedia content about a company and its products or services at their fingertips. As more and more companies use an intranet (internal Web sites) to transmit their webcasts, the reliability of the service greatly improves. The Yankee Group, a research firm, projected that companies will spend up to $5.9 billion on online videos by 2005 (Vonder Haar, 2002). The primary function of online videos is for advertising and marketing purposes (53%). Corporate communications, employee training, and sales training are other business functions of online videos.

Financial services companies also use webcasting to communicate with investment advisors. For example, Safeco, a Seattle-based mutual fund, produces 40 webcasts a year for 5,000 investment advisors who recommend Safeco funds and portfolios to clients. These webcasts feature interviews with portfolio fund managers every quarter to explain the different Safeco funds' investment strategies (Vonder Haar, 2002). Webcasts can become an important customer relation management tool that will substitute for meetings and conferences that are complex to arrange and cost much more.

Webcasting is also a means for international nonprofit organizations to disseminate information to developing countries. For example, the World Bank Institute recently launched B-SPAN, an Internet-based broadcasting initiative. B-SPAN will provide regular broadcasts of World Bank events through webcasts over the Internet. The webcasts feature some of the world's leading experts and practitioners in the financial, poverty, health, education, legal, environmental, and energy fields on the latest developments in their sectors.

To individuals who would like to create media contents and establish their own voices in the society, webcasting is an ideal medium. Not only is there no license requirement for the content provider, the cost of distribution to a geographically diverse audience is much lower than any traditional mass media, making the dissemination of media content for special interest and small audiences affordable. In addition, there is no censorship of media content on the Web (with the exception of some countries such as China) and the reach of the Internet can be global. Dissidents and minorities will find the Web the best place to broadcast views suppressed by the mainstream media.

Webcasters have four strategic business revenue sources:

(1) The subscription and pay per service model, which provides access to previously aired content or live content;
(2) The broadcast advertising sponsorship broadcast model, which offers companion content to supplement current on-air programming and displays advertising during the webcast;
(3) The e-commerce model, which sells products and complementary items on the Web; and
(4) The syndication model, which creates and distributes original digital media content to other webcasters.

Webcasters must choose one or a combination of revenue sources above to sustain or make profits from their webcasts.

EXAMPLES OF WEBCASTING

American Broadcasting Company (ABC), one of the big three U.S. broadcast TV networks, is webcasting part of its news content on its ABCNEWS.com site. ABCNEWS.com distributes both daily TV news content and past ABC News coverage on the Web via Virage's Internet Video Application Platform and Syndication Manager. It indexes and publishes searchable video clips from *Nightline* and *World News Tonight* and on ABCNEWS.com. Its news clips are also distributed to affiliates, news agencies, schools, and any other applicable Web sites (*Digital TV,* 2001). Its business model is primary a subscription model. Access to the basic Web page is free. However, additional news video and interview clips are served as premium content, which requires a monthly subscription fee of $4.95. It also partners with Real Network's RealOne SuperPass package as one of the program services offered to the package subscribers.

Real Network's RealOne SuperPass (Figure 1) is an example of a successful webcasting business model using subscription as the source of income (Schlender, 2002). According to Lisa Amore, the spokesperson for Real Networks, it has attracted 750,000 subscribers at the time of writing (personal communication, August 9, 2002). Its service is similar to a cable TV system's carriage and charges customers by subscription. The basic level subscribers pay a monthly subscription fee of $9.95 to receive 15 channels of news, sports, popular culture, and weather; 51 commercial-free radio stations; and an additional $10 to get legal digital music download service. The RealOne service provides a multimedia user interface for the Web. In one window, it delivers exclusive streaming audio and video content from its content providers such as Cable News Network (CNN) and Fox Sports; in adjacent windows, RealOne displays information of companion Web sites and other explanatory materials. The content providers share the revenue of the subscription with Real Network.

A webcasting service based on the broadcast advertising sponsorship model is OurMaine.com (Figure 2). This webcasting is a collaborative effort between WPXT-TV of Portland and WPME-TV. Visitors to the site can open streaming video clips with the news of the day. Some streaming content is picked up from other Fox Network affiliates to supplement locally produced materials. The station sends its video via FTP to its Web host. Page views at the site exceed 500,000 per month and the station

Figure 1: Real One SuperPass.

Figure 2: OurMaine.com.

Figure 3: ExtraNet TV.

supports the webcast through sponsorship from advertisers (*Digital TV,* 2001).

ExtraNetTV (http://www.extranettv.com/) uses an e-commerce model to provide more than 20 video shows on demand free to users with an online store of over 3000 gift items (Figure 3). The shows available to visitors range from comedy, horror/science-fiction, popular interest, technology, to health. Each show has one or more episodes and visitors can choose to view any version depending on their own connection speed.

WebFN is a financial news network that is a joint venture between Chicago-based Weigel Broadcasting and New York-based Bridge Information systems. Its webcasting works on syndication as the primary source of revenue (although it currently uses the broadcast model while it builds its syndication business). At the time of writing, it streams 12 hours of live video and data each business day over the Web and also on two Midwest TV stations. Programs are produced for use simultaneously on the Web and on television. The WebFN.com Web site also offers around-the-clock video on-demand service (Figure 4). Its trademarked "Viewcaster" presents a streaming video window and below it a window with interactive charts and graphs that are updated according to the in-stream news content. The webcast is presented in a "program wheel" format, rotating the featured segments at regular intervals during the hour plus five-program sector reports—"Markets in a Minute," "CEO:FYI," "Ask an Expert," "Bull Session," and "WebFN University." WebFN is working on syndicating their content to other financial Web sites such as Fidelity. They are also developing content partners in Europe and Asia in order to eventually offer a 4-hour live global webcast (*Digital TV,* 2001).

TYPES OF WEBCASTING

There are three types of webcasting based on the technology that webcasters use to deliver the content or information to the Internet audience: (1) push, (2) on-demand, and (3) live streaming (Miles, 1998). Webcasting can be streamed live or be downloaded and stored on the server for later retrieval by the users. Table 1 compares the differences between the three types of webcasting technologies based on consumers' effort and revenue sources.

Push Technology

Push technologies are computer programs that deliver the media content or information to the audience's computer screen automatically without specific request each time. The information may pop up as an alert, wallpaper, or screensaver on a person's computer, as electronic program guides on a TV screen, or as other displays on mobile devices or cellular phones. How the webcaster knows what content to push to the computer screen of the consumers is based on some level of intelligence such as the customer's needs or interests, previous information

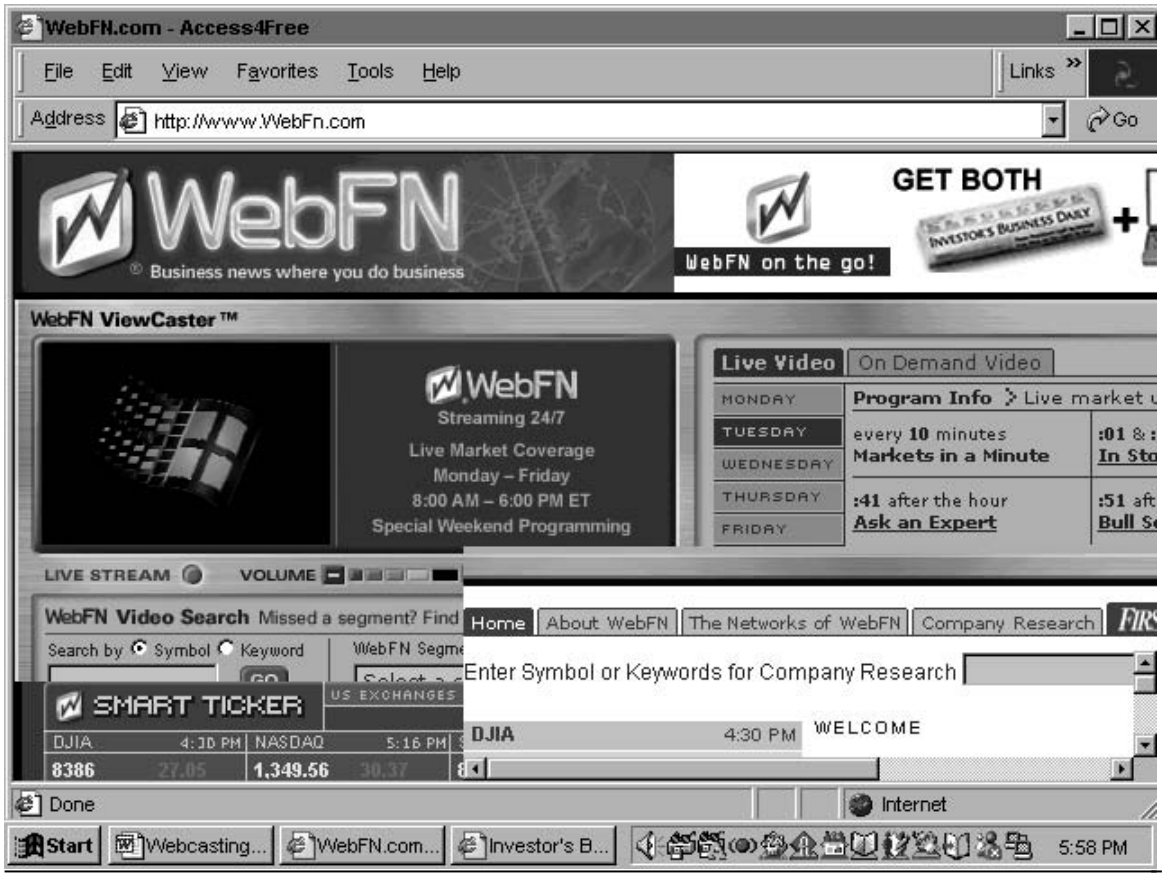


Figure 4: WebFN.

requests, or an affiliation or portal site registration such as Yahoo!.

One of the earliest companies that successfully employ the push technology on the Internet for webcasting was PointCast. Founded in 1992 to deliver news and other information over Internet connections, PointCast's flagship product PointCast Network sent customized news to users' desktops. PointCast Network was free to consumers and supported by advertising. To use it, one needed to download the PointCast client program, which was available from PointCast's Web site and other affiliated sites. The user provided preferences for customized information, which was delivered by Pointcast as screensavers. The company was acquired by EntryPoint in 1999 and is now defunct, its push service discontinued shortly after the acquisition. The pioneering push technology and proprietary software that Pointcast used encountered many technical difficulties such as slow service to customers, causing traffic jams on corporate networks. Other Web sites such as Yahoo! and Excite offered similar customized information services only when consumers used those sites. Consumers preferred this model over Pointcast's push model, leading to the termination of Pointcast's service (Bicknell, 2000). The failure of Pointcast does not mean that push technology is not a successful model of webcasting; it underscores the importance of bandwidth limitation in the use of pushing technology. Electronic mail software companies successfully and widely use the push technology to disseminate information to their users. Software companies push product updates and downloads to the users such as the latest version of McAfee antivirus programs, Eudora, Internet Explorer,

Table 1 Comparison of Three Types of Webcasting Technologies

	Push	On-demand	Live streaming
Consumers' effort	None/Automatic	High (must locate where the content is available and know what content to ask for)	Moderate (must know the webcast schedule)
Business model	Advertising and e-commerce	E-commerce, pay per service, and subscription	Advertising, pay per service, subscription and syndication.

and Netscape. Audio, graphics, and Web pages are sent automatically to the users' e-mail boxes.

The push and pull technologies work together to provide consumers effortless access to information of interest to them. Pull technologies use software to pull information for the consumer from the Internet. Consumers can set up their own selections, or the intelligent software agents can search out information that consumers would be interested in. Amazon.com, for example, has pull technologies that remembers the products consumers have purchased in the prior visits and pull information of similar products to the user on the next visit.

The greatest benefits of using the push technologies is that consumers can download programs, receive news updates, and interact with the webcaster in different formats such as audio, video, text, and graphics without actively searching for the information. It brings "intelligence and efficiency to the distribution of all kinds of information, giving webcasters more control over what users see and when" (Miles, 1998, p. 24). Push technologies provide effortless reception of materials with little consumer knowledge requirement and save consumers the time to find information on the Internet.

The latest development in push technologies is to become more interactive and automatic. Webcasters increasingly provide information based on users' past behaviors and their response to previous information. Editors are seldom used to sort or filter content. Some inherent problems of push technology surround how much information should be provided to the audience and the potential invasion of Internet users' privacy by pulling intelligence from users' past choice and preferences without their explicit consent. Users are increasingly intolerant of unsolicited materials that webcasters send without permission. They will tune out these materials and the effectiveness of pushing materials is thus greatly diminished.

On-Demand

On-demand webcasting refers to the webcasting of content based on the request of the consumer at the time of use. The webcaster supplies the media contents in a catalog or playlist for consumers to choose. The term on-demand refers to the ability of the user to control the scheduling and appearance of the webcast. The audience, instead of the webcaster, chooses when and what to view. In many on-demand webcasts, additional features are provided, such as instant replay; no waiting for rewind or fast forward during live events; interactive devices such as question and answer, chat rooms, product or service order forms, multiple camera angles, and zooming in and out the picture. On-demand webcasting's greatest attractions are the convenience to the consumer and the audience's ability to retrieve information or content that has been missed. There are both business and consumer uses of on-demand webcasting. Business and governments are already making use of on-demand webcasting to provide training and product information for employees, suppliers, and customers. Intel, for example, frequently uses webcasts to provide live seminars to prospective customers.

On-demand webcasting can be downloadable or show the content in real time. If the webcast is downloadable, compressed files are downloaded to the user's computer and a media player is used to play back the files (also called HTTP streaming). On-demand can also use real-time streaming to broadcast the content directly from a server as selected by the user so that the user can instantly watch or listen to it. No file storage or cache is involved in streaming (Real Networks, 2002). Users cannot store, copy, or retransmit the content because the stream never touched the local drive of the user and the data are discarded after playback. Hence the content owners can prevent piracy of their content. In addition, on-demand can extend webcasters' reach, increase the value of their media contents over a long period of time, and save bandwidth as well because not everyone connects at the same time (Mack, 2002).

Live Streaming

Streaming is an Internet data transfer technique that does not need to wait for the whole file to download to a user's computer before playback begins (Mack, 2002). It involves three different types of software: (1) an *encoder,* which converts an audio or video signal from an analog format to a digital format and compresses the digital files for transmission over the Internet; (2) a *server,* which delivers data streams to audience members; and (3) a *player,* which viewers use to watch or listen to the streaming media. The content of webcasts using streaming technologies can either be live or on-demand. In live streaming webcasts, the broadcast schedule is fixed so that time-sensitive contents can be delivered instantaneously to users. Users must follow the schedule to receive the content. Live webcasts require multiple streaming servers to reduce the load of each server and to minimize the chance of crashing a server by too many users tuning in at the same time.

In a live streaming webcast, users can also interact with the event as it is happening such as clicking on a picture of a car to view its details or change its color. Although many refer "streaming" to the retransmission or repurposing of traditional broadcast content such as television and radio signals on the Web, streaming is not limited to retransmission. It can also be original content transmission with intent to reach the public. Webcasting using streaming technology can pick up viewers or listeners from areas around the world.

The ability to squeeze the audio and video into a stream is the basis of streaming technologies. These technologies use software to compress the signals for transmission and decompress the signals for display on the viewer's screen in the correct order continuously so that the viewer can view the content almost instantly. It is very important that the signals are reassembled in the correct order during the decompression process. Any gaps in between will make speech and video incomprehensible. For successful webcasting using streaming technology, the webcaster needs to secure a powerful high-speed server to deliver the signal, and they need enough bandwidth to enable multiple users to view the program.

Multicasting can be a method to save on bandwidth costs for webcasters. It is the ability to take one signal and send it to lots of people through a network or over the Internet. The one signal locates a device (e.g., a router) that sends the signal to a number of computers or television

Table 2 U.S. Webcasting Audience Profile

	Ever tuned to a webcast	Tuned to a webcast in the past month	Tuned to a webcast in the past week
In U.S. Population	80 millions	40 millions	20 millions
% age 12+	35%	17%	9%
% Internet users	48%	24%	13%
% broadband users	26%	30%	34%
Male/female	50%/50%	56%/44%	62%/38%
Under age 35	52%	56%	60%
Time spent on Internet/day	1 hr 42 minutes	2 hr 16 minutes	2 hr 49 minutes

Source: Arbitron/Edison Media Research 8 Study, 2002.

sets. Multicasting is much cheaper than unicasting, which assigns one stream to each viewer or listener. When multiple users log on to the site, the unicasting webcaster will need to send multiple streams that take up large amounts of bandwidth. One drawback of multicasting is that the webcaster needs to enable multicast in routers and cannot use automatic rate changing to accommodate the different connection speeds of users (Austerberry, 2002).

Among the three technologies, on-demand and live streaming are currently the most well-known webcasting mode and most popularly used in consumer webcasting of entertainment content. According to the latest Arbitron/Edison Internet 8 study (Rose & Rosin, 2002), nearly 80 million Americans aged 12 or above have either listened or watched a webcast online, with about 7.2 million having tuned to one in the past week. Surprisingly, webcasting use is not limited to young people. About 44% of webcast users are older than 35 years of age. Table 2 provides a profile of webcast audiences in the United States.

Three Levels of Webcasting

Webcasting can be differentiated into three levels based on the degree of sophistication in the webcasting technologies used during the webcast. An example of low-end webcasting is pushing information by e-mails. E-mail campaigns targeted at customers, suppliers, and business associates that have actually requested information are a proper use of the low-end webcasting to market products and build customer relations. The e-mails can include Web page links and audio and video files. Those unsolicited mass mailings via e-mail, usually called spams, do not discriminate the identity of the recipients and are sent from sources unknown to the audience. These spammers are under the scrutiny of state antispam laws and subject to prosecution.

The mid-range webcasting is the placing of video or audio content on a Web site and providing customers and associates with 24 hour access to current events and information about a company's products or services. The webcasting can be accompanied with features that enhance the video and audio experience. These features range from search engines or directories that help visitors to find specific information, to captioned and cued slides or diagrams that can be displayed along with the audio or video.

High-end webcasting applications may be either push or streaming, or a combination of the two. High-end webcasting is similar to traditional broadcasting because of its expected large audience. Nonetheless, it differs from traditional broadcasting by its video library accessible on demand, 24 hr a day. Businesses use high-end webcasting to disseminate information to remote office locations or to reach prospective customers or investors. The entertainment industry is also using high-end webcasting to enhance the viewing experience with the so-called "enhanced TV" and to generate additional revenue source with file downloading services for a fee and purchase of video content online. Webcasters need to lease high-speed telephone lines, satellite delivery, and other ways of transmitting the live webcast signal to the Internet connection. They also need to purchase sufficient bandwidth that will meet the demand of the large number of people logging onto the server at the same time.

TECHNICAL STANDARDS AND PROTOCOLS OF WEBCASTING

The Internet is a collection of computer networks that are interconnected and communicate with each other based on a common protocol called TCP/IP (transmission control protocol/Internet protocol). Protocols define the way in which one hardware or software component interacts with another with respect to specific functionality. As webcasts are shown to different people with different computers and Web display devices, protocols and standards are needed to enable communication across networks. Protocols become standards when every webcaster uses them. Several agencies set the standards for the webcasting industry. The governing standards agency for the Internet is the IETF (Internet Engineering Task Force). Other organizations such as the ITU (International Telecommunication Union), MPEG (Motion Pictures Expert Group), and W3C (World Wide Web Consortium) also create standards and their standards have become more important for the industry (Miles, 1998). Figure 5 illustrates the relationships among the webcasting protocols that work collaboratively to send a transmission to a user.

IETF (Internet Engineering Task Force) Standards

The IETF is an international association of network designers, operators, vendors, and researchers concerned

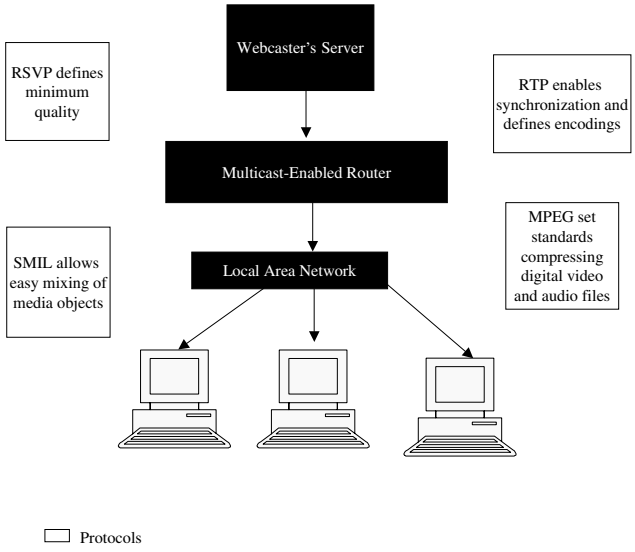


Figure 5: Webcasting protocols in multicasting.

with the evolution of the Internet. Four webcasting standards have been adopted by IETF:

IP multicast;
Reservation protocol (RSVP);
Real-time transport protocol (RTP) and real-time control protocol (RTCP); and
Reliable multicast protocols.

IP Multicast

IETF identified three ways to transmit from a source to multiple recipients on the Internet:

(1) unicasting—point-to-point transmission;
(2) broadcasting—one-to-all transmission; and
(3) IP multicasting—one copy is sent to a group address.

Unlike unicasting, IP multicasting allows small or large amounts of digital information to be sent to large audiences. Only group members that should receive the webcasts will actually receive the programs and only one copy of the information is needed to first reach a group address, then routed to individual recipients to allow efficient delivery of digital information.

Reservation Protocol (RSVP)

To ensure the quality of the webcast, which includes integrity, end-to-end predictability, and efficient bandwidth utilization of data transmission, it is necessary to specify the minimum quality standard. RSVP, the reservation protocol, is one kind of quality standards that enhances the current Internet architecture with support requests for a specific quality of service from the network for particular data streams or flows. This protocol is designed to allocate network resources appropriately for the requirements of the data being sent. To optimize transmission for particular types of data such as audio and video, RSVP defines the network traffic class and is used to control both quality of service and resource management for unicast and multicast sessions.

Real-Time Transport Protocol (RTP) and Real-Time Control Protocol (RTCP)

Because audio and video webcasts transmitted over the Internet can be lost or experience variable delays, RTP is a protocol intended to enable synchronization and recovery from loss or delays. RTP also defines a format for different audio and video encodings to promote interoperability among different computer platforms, operating systems, and application software products. By having specific data fields that contain timestamp and sequence information, the receiving computer can use these fields to reconstruct the time-specific properties of the RTP data streams. A related protocol is RTCP, which checks the status of a webcast from time to time. Using the RTCP, sender and receiver reports are transmitted from time to time so that applications using RTP can get RTCP reports on how well RTP data are being delivered.

Reliable Multicast Protocols

Reliable multicast protocols aim to offer 100% data integrity over a network when needed. Sometimes we don't need complete data integrity such as when watching a movie because human eyes and ears can tolerate and compensate for minor loss or interference in sound and pictures. However, for transmitting databases or software, no loss can be allowed, and it is necessary to use software that is built in with a reliable multicast protocol.

ISO (International Standards Organization) Standards

The International Standards Organization (ISO) is a worldwide organization of national standards organizations from over 100 countries to promote the development of standardization and related activities in the world. Its goals are to facilitate the exchange of goods and services and to develop international cooperation in intellectual, scientific, technological, and economic activities. MPEG is a group of people who met at ISO to generate standards for digital video and audio compression. MPEG-1 audio compression standards is composed of three coding levels: Layer-1 (MP1), Layer-2 (MP2), and Layer-3 (MP3). Each layer level is a higher compression ratio at equal audio quality. To reproduce CD quality audio, Layer-1 requires 384 Kbps, while Layer-3 only requires 112 Kbps. MP3 has now become the most popular standard for digital encoding and transmission of audio and video. MPEG-4 is the latest standard initiated by the ISO. Different types of multimedia can be combined for different presentations using MPEG-4 (Mack, 2002).

ITU (International Telecommunications Union) Standards

ITU (International Telecommunications Union) is an international organization where government and the private sector work together to coordinate global telecommunications network and services. Headquartered in Geneva, Switzerland, the ITU has played an important role in standardizing the videoconference industry with

the H.320 suite of audio/video compression standards. The ITU has also provided the standards for multipoint document conferencing with its H.323 standards, which standardizes conferencing over packet-switched networks such as the Ethernet. It is also currently working on standards for electronic program guides (EPG) that will affect webcasting and the listing of webcast events on the Internet and television. The T120 standard contains a series of communication and application protocols that provide support for real-time, multipoint data communications. Over 100 webcast and streaming media suppliers such as Microsoft and Cisco Systems have already committed to implementing T120-based products.

World Wide Web Consortium (W3C) Standards

The World Wide Web Consortium (W3C) is an international industry consortium founded in 1994 to develop common protocols for the Web's evolution. The consortium's current efforts on synchronized multimedia and extensible markup language (XML) are the most significant to the webcasting industry.

The synchronized multimedia project is to establish standards that enable synchronization of different media (text, graphics, audio, video) so that the presentation can be shown in a coordinated way. As a result of the project, W3C proposed a new markup language for use on the Web called synchronized multimedia integration language (SMIL). This language was designed to allow the easy mixing of simple media objects in different formats. The coding would use simple tags to designate elements on a Web page. It will make it easier for people to design and add webcasting elements to their Web pages. The use of the language will increase the accessibility of the sites through standardization of media objects display and it will also increase the accessibility of audio-enabled Web sites for the visually impaired.

XML is an advancement from HTML (hypertext markup language), which has been the basis for building Web pages. It is a much more flexible language than HTML and allows designers to define their own customized markup language, enabling the use of standard generalized markup language (SGML) on the Web, which can define, identify, and use the structure, style, and content of documents.

Proprietary Protocols

Apart from the standards-setting agencies, several webcasting protocols are proprietary to a software vendor, but they are eventually submitted to the standards-setting agencies for consideration to become a common standard for the entire industry. Two such protocols are real-time streaming protocol (RTSP) and advanced streaming format (ASF).

Real Networks, Netscape Communications, and Columbia University developed the real-time streaming protocol jointly. RTSP is an application-level protocol for control over the delivery of data with real-time properties. The protocol is designed so that delivery of both live data feeds and stored content can be brought under the control of the webcaster. RTSP has been submitted to the IETF for standards consideration.

Advanced streaming format has been introduced by Microsoft to define the storage format for streaming media. ASF is an open standard file format in which the tools, servers, and clients of multimedia vendors store, stream, and present multimedia content in the same file, instead of as separate audio, text, graphic, and video files. Microsoft has submitted ASF for standards consideration with the ISO and IETF.

STATE OF THE RADIO WEBCASTING INDUSTRY

Radio is the most common form of webcast tuned in by U.S. Internet users mostly during work hours (Measurecast, 2002). The latest Arbitron/Edison Internet 8 study (Rose & Rosin, 2002) shows that 25% of Americans aged 12 or above have listened to a radio station on the Internet, although weekly listening was composed of only 4% of online radio listeners. Only 45% of the online radio listeners listened to local radio stations most often. The rest most often listened to stations from other parts of (41%) or outside the United States (9%).

The audience size of individual radio webcast is still small. Among the top 10 radio webcasts, none of them have a monthly cumulative measured audience (CUME) of more than 400,000 listeners, according to Measurecast, Inc., the company that provides next-day audience reports for advertisers and Internet broadcasters (http://www.measurecast.com). MusicMatch, which provides an Internet-only radio webcast, charging a subscription of $3.33 a month, topped with a CUME radio webcast audience of 365,783 in September 2002. Virgin Radio of London, which has the highest CUME, only had a CUME of 264,788 in September 2002. The U.S.-based radio station that had the highest CUME during that same period was WQXR-FM, a New York-based classical music station. Its monthly CUME is only 83,550.

According to BRS Media's study in April 2000, 9,321 radio stations had Web sites, of which 5,945 were U.S./Canadian radio stations. The top 10 features of radio Web sites, according to Arbitron/Edison Media Research, are DJ info/pictures, community events, links to advertisers, cool links, station information, contest entry forms, program schedules, concert information, e-mail contact, and station listening link (Gunzerath, 2000).

Levi's (2000) white paper presented to the NAB radio show reported that 37% of all radio stations offer streamed audio, and there are a growing number of Internet radio stations that only operate exclusively on the Web for Internet audiences. Because most Web radio originates from terrestrial broadcast radio stations, the resulting Web radio stations base much of their content on their terrestrial broadcast radio station counterparts.

STATE OF THE TELEVISION WEBCASTING INDUSTRY

Compared to the radio webcasting industry, the TV webcasting industry received much less attention from the

press. Nitschke (1999) conducted a survey regarding TV stations' Internet operations, which found about 70% maintain their own Internet operations with an in-house staff. Among those 37 stations who answered the survey, about 30% have streaming media services providing audio and video files. Only 14% have chat rooms and 27% have a listserv for e-mail broadcast. Some TV companies, such as NBC and the Hearst-Argyle television station group, subcontract the webcasting operation to special webcasting service companies such as the Internet Broadcasting Systems (IBS). IBS develops and hosts the sites, hires and manages the Web editorial staff, and builds revenues for the stations. All its sites serve archived and streaming videos and have a similar design to that of members of the IBS network. Chan-Olmsted and Ha's (in press) recent survey of TV station managers show that customer relation management and collection of audience intelligence are the main goals for TV stations to establish their online presence. The revenue stream that most broadcast stations expect from their webcast is advertising (76%). Barter (24%) and e-commerce (19%) are a distant second and third.

Transferring their domination on television, the major national TV networks' webcasts are the most popular among TV webcasts. CNN, MSNBC, and CBS News are the forerunners in TV webcasts, according to the DFC Intelligence Research Study (Miles & Sakai, 2001). With the high demand of bandwidth for video streaming, it is not surprising that TV webcasting is used much less than radio webcasting. Arbitron/Edison's Internet 8 study (Rose & Rosin, 2002) revealed that only 7% of Internet users watch a video webcast, which is three and a half times less than radio webcast numbers. The major obstacle of video webcast is the requirement of a high-speed Internet connection for Internet users to watch a video webcast.

MAJOR PLAYERS IN THE WEBCASTING INDUSTRY

Seattle-based software companies such as Microsoft and Real Networks are important players in the webcasting industry. According to the Arbitron/Edison's Internet 8 study (Rose & Rosin, 2002), the top four streaming media service providers to consumers are Real Network's Real.com, Yahoo Radio, MSN Music, and Radio@AOL. Microsoft's Windows Media Player and Real Network's RealPlayer are the most frequently downloaded streaming video players. According to the January 2002 Nielsen/Netratings, Real Player is used by 32.9 million U.S. households while Windows Media Player is used by 14.2 million households. Apple Quicktime, as a latecomer to the webcasting industry, is now a popular multimedia player among people under 21 with 8.1 million home users (Miles & Sakai, 2001; Olsen, 2002). Real Networks recently launched the Universal RealOne Player, a multiplatform media player that can play and cache all media types including Apple Quicktime and Windows Media.

The International Webcasting Association (IWA) is the largest worldwide nonprofit trade organization representing companies, organizations, and individuals active or interested in the delivery of multimedia (http://www.webcasters.org). The IWA is headquartered in Washington, DC, and serves members throughout the United States, Europe, Asia, Canada, and Australia.

PROBLEMS AND ISSUES IN WEBCASTING

The webcasting industry is facing several problems. The first problem is the cost of the digital content and copyright issues. They have to protect the content from being stolen or reproduced by other webcasters and users while securing contents with reasonable fees. Secondly, webcasters compete for audiences with offline media such as television and radio. Thirdly, webcasting quality and delivery efficiency depend on the degree of broadband Internet access penetration of the market. Fourthly, webcasting content development needs to take into account the bandwidth requirement and downloading time to users. Lastly, the regulations of the Internet and on the freedom of speech can greatly affect webcast contents.

Cost of Digital Content and Copyright Issues

The cost of digital content and copyright protection are currently the larger issues concerning webcasters especially Web radio, which broadcasts music on the Internet. The report of the Copyright Arbitration Royalty Panel (CARP) recognizes the sound recording and musical work performance rights of the sound recording industry and recommends that webcasters pay the high royalties charged by the sound recording companies. Specifically, webcasters must pay for each performance by the number of listeners. After the Library of Congress revised the royalty rates, webcasting companies with or without an offline radio station will pay the same rate of $0.07 per person. This rate is nearly 10 times higher than the rate suggested by webcasters and approximately 35% of the royalty fee requested by the Recording Industry Association of America (RIAA). Both the National Association of Broadcasters and the International Webcasting Association protested against the CARP recommendations (IWA). Many Web radio stations have shut down to protest the decision and avoid being fined. In a survey of readers of *Streaming Magazine,* who mostly are webcasters and streaming technology suppliers, 80% think that the CARP-suggested royalty rates will kill the Internet radio business (Jeffrey, 2002).

The Small Webcaster Settlement Act (HR5469) was introduced and passed in the U.S. Congress recently in November 2002 to help small webcasters. The bill offers a lower royalty rate to small commercial webcasters by only charging 8% of the gross revenue or 5% of the webcasters' expenses for the royalty fee for the retroactive period before December 2002. Now webcasters with less than US$50,000 in gross revenue will pay a minimum annual fee of $2,000. Other webcasters will pay a 10% royalty fee out of the gross revenue for the first $250,000 and 12% of the gross revenue when the revenue exceeds $250,000; or 7% of the operating expenses, whichever is greater. However, the bill does not apply to college webcasters so

there are still some issues to settle (full text of the bill can be found at http://frwebgate.access.gpo.gov/cgi-bin/getdoc.cgi?dbname=107_cong_bills&docid=f:h5469eas.txt.pdf).

Because of the ease to duplicate in identical quality as the original and to retrieve and store digital content on the Web, webcasters must protect their own copyright for original content. The U.S. Congress enacted the Digital Performance Right in Sound Recordings Act of 1995. This act only applies to digital audio retransmission and requires webcasters to obtain a performance license from the owners of the sound recording rights. In addition, the Digital Millennium Copyright Act (DMCA), written in 1998 under the treaties by the United Nation's agency World Intellectual Property Organization, provides for a simplified but statutory licensing system for digital performance of sound recordings on the Internet and via satellite. Its provisions include a programming restriction called the "Sound Recording Performance Complement" (SRPC). The restriction includes no more than three songs from a particular album, no more than four songs by a particular artist, and no more than three consecutively in a three-hour period and no advance song or artist playlist announcement may be published. The current controversy over the Copyright Arbitration Royalty Panel's report and recommendation is a result of that Act.

Competition for Audience with Offline Media

Although some have argued that webcasting serves a totally different type of audience from other traditional or offline media such as television and radio with its interactive capabilities and niche programming, it is still a medium that competes with offline media for both advertising dollars and audiences. The growth of webcasting will result in more audience fragmentation, as thousands of webcasts are available on the Internet. Nevertheless, webcasting also brings in new audiences—audiences at work and audiences not served by niche programming or limited by local programming in the past. In order to win the competition for audiences, both webcast media and other offline media must work even harder to understand and discover an audience's unmet needs and provide the service that meets those needs.

On-demand webcasters may have a tough time in competing with digital cable TV's videos-on-demand service, which offers consumers the choice to watch programs on TV in digital quality whenever they want. The interactive services provided on digital cable can be comparable to the interactive devices in a webcast. In addition, the increasing penetration of personal video recorders (PVRs) that allow live recording and skipping of commercials on television is another potential competitor to on-demand webcasting to consumers. Consumers can shift their TV viewing schedule and select the content they want easily with PVRs, making webcasting a less appealing viewing option. The advent of satellite radio may reduce the attractiveness of Web radio because satellite radio can provide radio service with digital sound quality on the road on a national basis without commercial interruptions. The major selling point of webcasting is its transcendence of geographic barriers because cable, satellite radio, and PVRs are all tied to either a local or a domestic market.

Broadband Access and Speed of File Delivery

High broadband market penetration is a prerequisite for the growth in the use of webcasting because the connection speed of dial-up modems is too slow to provide acceptable audio and video quality for entertainment use. Currently, about 30% of Internet users in the United States are broadband users (i.e., 33.6 million) according to Nielsen/Netratings (January 2003). To users without broadband access, viewing video is not only inefficient but a frustrating experience because the gaps between streaming data transmission created garbled and incomprehensible images and audio sounds. The increase in broadband use results in an increase in webcasting usage. Nielsen estimated that about 12.7 million broadband Internet surfers consumed streaming media content at home. The bandwidth requirement for a large number of audiences in webcasting is also a big hurdle for many webcasters when the public Internet system is used. For example, when Madonna's live concert in London was webcasted in November 2000, nine million streams were served to the Internet audience. MSN produced the show to generate publicity. However, the webcast cost MSN so much that MSN could not afford to pay Madonna the concert rights.

Development of Webcasting Content

The development of webcasting content must take into consideration the bandwidth requirement and transmission quality. Videos, for example, take up a large amount of bandwidth, and should therefore be used very selectively. Short video clips and interactive features are much more effective than full-length video in webcasting. Animation and showing images in a sequence ensure the reception quality to the user because the loading time for images and visuals is much less than a video. They can be as effective as a video in many instances.

Navigation design is another important issue in developing webcasting content. Webcasters need to organize the content offering into separate categories and display the approximate playing time by the modem speed of the user. A user-friendly menu will enable the user to quickly locate the webcast content to be played and control the pace of the webcast. If the contents have been used in other offline media, a reference to the original aired dates will be of high reference value to the users. Digital copyright management is another important aspect in webcast content development.

Regulatory Issues in Webcasting

Apart from copyright protection, there are several regulatory issues pertinent to webcasting: (1) extension of the right of publicity online (state laws prohibiting the unauthorized taking of an individual's name, likeness, voice or other elements and using them for commercial purposes); (2) invasion of privacy by the Internet's ability to collect data and customize information; (3) libels in bulletin boards and chat rooms; and (4) freedom of speech

on the Internet and censorship issues. The Communication Decency Act of 1996 was ruled unconstitutional by the Supreme Court, which established that the Internet enjoys the same First Amendment rights as other print media formats. Essentially, the access and cost to use Internet services determine how many will be able to use webcast.

WEBCASTING AROUND THE GLOBE

The United States is taking the lead in the webcasting market development by promoting private investment for universal Internet access. Europe is lagging behind in webcasting market development because of the charge per use rate structure of local phone service and the high price of leased lines and dedicated circuits. Asia is also generally underdeveloped in consumer webcasting because of limited press freedom in most Asian countries. Nevertheless, the high broadband adoption rate in the Four Little Dragons in Asia (Hong Kong, Singapore, South Korea, and Taiwan) and the vibrant business environment foster a fertile ground for business-to-business webcasting. B2BCast is a webcasting company based in Asia providing business-to-business webcasting. It focuses on providing business executives on-demand access to business events and conferences 24 hr a day, such as the Asian Corporate Branding Symposium. The Hong Kong Trade Development Council's web site features business webcasts with eight different channels (http://www.tdctrade.com). Consumer webcasting is blossoming in Taiwan with the encouragement of the government on broadband development. For instance, HiNet, a Taiwan broadband Internet service provider (http://www.hichannel.hinet.net), provides its subscribers live webcasts of swimsuit model shows every evening and celebrity interviews. Latin America's webcasting industry is hampered by its low Internet usage rate. Brazil, Argentina, Chile, and Mexico are the most highly developed Internet markets in Latin America. However, less than 5% of their population uses the Internet (Cyberatlas, 2002).

CONCLUSION

Webcasting is still an infant industry with many small entrepreneurs and large enterprises experimenting their way. A liberal regulatory climate and broadband Internet access adoption are key environmental factors to foster the growth of this industry. Either willingly or unwillingly, traditional broadcast media such as radio and television have participated in the brave new world of webcasting. Does webcasting pose a threat to traditional broadcast media or will its presence further strengthen the value of the traditional media? Media history repeatedly shows that each medium will modify itself to adapt to the changes in the environment, find its own niche, and survive the threat of new media, unless the new media can completely substitute for the functions of the traditional medium such as the replacement of vinyl records by compact discs and cassettes. Just as television was unable to displace radio, and radio was unable to displace newspapers, the Web will not displace television or radio. Instead, audiences are given more choices with webcasting. Delivering of multimedia content is much easier now than before with the different webcasting technologies and advancement in webcasting protocols. By integrating the latest technology on the Web and the traditional media content, broadcasters may find higher value in their media content and a much larger audience in the workplace and home and around the world. The interactive capability of the Web opens the door for a variety of revenue streams for the most creative broadcasters/webcasters who best serve the needs of consumers and know how to maximize the value of their content.

GLOSSARY

CUME (cumulative measured audience) The number of different or unduplicated homes/people exposed to a program at least once across a stated period of time.

Extranet An external Web site created by an organization to provide information and support services with restricted access, such as passwords or security codes, to its customers, clients, suppliers, or members.

FTP (file transfer protocol) The Internet application of moving files across the Internet using the TCP/IP protocol by either uploading a file to a computer server or downloading a file to a user's local computer drive.

Intranet An internal Web site created by an organization to provide information and communication for its employees, its purpose being primarily for business usage but also to be used as an internal corporate communication tool.

NAB (National Association of Broadcasters) A broadcast industry association representing the interests of free, over-the-air radio and television broadcasters.

Pull The process of using software to find information on the Web for the consumer to view through a browser or e-mail; can be initiated by the consumer or automatically processed by the Web site server.

Push The delivery of information by a company to a user's computer on a regular basis, ranging from simply sending regular e-mails to providing customized information in multimedia format based on either the user's request or automatic intelligence agents that determine the user's tastes and preferences.

Real time The delivery of media content, data, audio, or video at almost the same moment it originates on the Web; equivalent to a "live" broadcast on the computer screen.

Repurpose The media content management strategy of using the same text, audio, or graphics content again in other media channels by the copyright owner of the content, the term originating from content created for one purpose (a TV newscast) being used for another (a webcast), which can be longer or shorter than the original content or exactly identical. Also the media content is still considered the same and under the protection of copyright law even though the distribution media have changed.

RTP (real-time transport protocol) An Internet protocol that provides a timestamp and sequence number to facilitate the data transport timing and to control the media server so that the video stream is served at the

correct rate for transmitting real-time data in webcasting.

Satellite radio National radio service supplying many radio channels via satellite. Users need to buy a satellite radio receiver to receive those satellite radio signals. Currently, service providers such as XM and Sirius provide satellite radio services in the United States on a subscription basis.

Streaming The technology of sending a continuous data signal through the Internet with special software; enables the user's computer to decode a signal as soon as it is received and play it almost immediately in the correct order. Unlike downloading, which requires the storage of the data in the user's local hard drive before playback, streaming data are not cached (stored) in the user's local computer drive and play back the data at almost the same time as they are transmitted.

CROSS REFERENCES

See *Copyright Law; Extranets; Intranets; Video Streaming.*

REFERENCES

Austerberry, D. (2002). *The technology of video and audio streaming.* Woburn, MA: Focal Press.

Bicknell, C. (2000, May 29). Pointcast coffin about to shut. *Wired.* Retrieved October 31, 2002, from http://www.wired.com/news/business/0,1367,35208,00.html

Chan-Olmsted, S., & Ha, L. (in press) Internet business models of broadcasters. *Journal of Broadcasting and Electronic Media.*

Cyberatlas (2002). *The world's online population.* Retrieved November 20, 2002, from http://cyberatlas.internet.com/big_picture/geographics/article/0,1323,5911_151151,00.html

Digital TV (2001, February). Webcasting. 45–50.

Gunzerath, D. (2000). *Radio and the Internet.* Retrieved May 13, 2002, from http://www.nab.org/Research/topic.asp#INTERNET

Ha, L., & Chan-Olmsted, S. (2001). Enhanced TV as brand extension: TV viewers' perception of enhanced TV features and TV commerce on broadcast networks' Web sites. *International Journal on Media Management, 3*(4), 202–212.

International Webcasting Association (IWA) (2002). Retrieved April 17, 2002, from http://www.webcasters.org/

Jeffrey, J. O. (2002, April). Will CARP kill Internet radio? *Streaming Magazine,* 25–26.

Levi, T. (2000, September 20). The adaptation of streaming media: Update 2000. Radio 2020: A sound vision of radio's future. White paper presented at the annual NAB Radio Show, San Francisco, CA.

Mack, S. (2002). *Streaming media bible.* New York: Wiley.

Measurecast, Inc. (2002). Measurecast reports Christian music format popular with Internet radio listeners. Retrieved August 28, 2002, from http://www.measurecast.com/news/pr/2002/pr20020827.html

Miles, P. (1998). *Internet World guide to webcasting: The complete guide to broadcasting on the Web.* New York: Wiley.

Miles, P., & Sakai, D. (2001). *Internet age broadcaster* (2nd ed.). Washington, DC: National Association of Broadcasters.

Nielsen NetRatings (2003, January). Broadband access grows 59 percent, while narrowband use declines, according to Nielsen/Netratings. Retrieved April 14, 2003, from http://netratings.com/pr/pr_030115.pdf

Nitschke, A. (1999). Station Internet activities report. Retrieved May 10, 2002, from http://www.nab.org/Research/Reports/TvstationInternetActivity.asp

Olsen, S. (2002, June 19). Apple: We told you QuickTime was #1! *CNET News.com.* Retrieved April 14, 2003, from http://zdnet.com.com/2100-1105-937379.html.

Real Networks (2002). Streaming media F.A.Q. Retrieved August 5, 2002, from http://www.realnetworks.com/resources/startingout/get_started_faq.html

Rose, B., & Robin, L. (2002). Internet 8: Advertising vs. subscription—Which streaming model will win? Arbitron Inc. and Edison Media Research. Retrieved August 7, 2002, from http://www.arbitron.com/home/content.stm

Schlender, B. (2002, March 4). The real deal. *Fortune,* 215–220.

Vonder Haar, S. V. (2002, July). Streaming media gets serious. *Baseline,* 83.

Web Content Management

Jian Qin, *Syracuse University*

INTRODUCTION

Web content management is a process in which content management solutions are used to create, store, publish, update, and repurpose content to be communicated through an organization's Web site. The content may appear in the form of HTML, XML, image, audio/video, plain text, or database. The management of these assets is achieved through using templates, workflow tracking features, publishing systems, and storage of content objects in database or file indexing systems. Web content is rapidly becoming one of the primary components of the competitive advantage for all types of organizations. Whether an organization's Web site can attract and retain users plays an important role in an organization's success, especially for e-businesses. The increasing importance of Web content demands better and more responsive methods and technologies that produce and provide access to it.

Web pages of the early days were created mainly through manual HTML coding and the content contained in these pages was static. As the number and types of files composing a Web site grew rapidly, manual coding and linking became time-consuming and error-prone. The situation could worsen when more than one person worked on the same Web project. The need for effective and dynamic publication of data and link creation on the Web prompted a new generation of the Web. The early dynamic generations of Web content primarily used programming interfaces and languages such as the common gateway interface (CGI) and Perl. Later, other programming languages, e.g. VBScript, JavaScript, and JavaServer Pages (JSP), became popular. By using computer programs and stylesheets, data stored in databases could be browsed or retrieved and displayed on the fly in a predefined style. Although dynamic presentation of data on the Web was accomplished by utilizing information technology, new challenges emerged. One such challenge was version control. In the case of a large team working on a Web site that contains hundreds or thousands of files, poor coordination and lack of version control can cause unnecessary damages to the work being done. Another problem is the workflow management. As maintaining a Web site becomes part of the daily life of an organization, tasks such as creating, editing, testing, approving, and publishing will involve many people at different levels and in different specialization areas. A lack of rules and procedures can create bottlenecks or blockages of the workflow. These two example problems demonstrate that creating and maintaining a Web site, particularly a large one, is no longer a simple matter of assembling a group of HTML pages and programs. More sophisticated systems are required for the production and delivery of Web content. These systems are expected to solve problems such as content creation, representation and organization, asset management, access management, production content delivery, version control, and scheduling.

Web content management becomes the term referring to a system that performs all these tasks, as well as the practice that uses the system to produce, deliver, and manage Web content. It is therefore used to mean both processes and the technology involved. This chapter will discuss Web content management primarily following the processes from content creation to delivery. Key technologies will be mentioned when necessary, but vendors for particular products will not since they are not the focus of this chapter.

WEB CONTENT LIFE CYCLE

Components in a Web content life cycle are given different names depending on how developers view and build the content management system. Latham (2002) proposes a six-phase life cycle for more general content management. They are creation/acquisition, review, aggregation/management, distribution, archiving, and destruction, all in the context of workflow and integration. IBM defines a Web content life cycle as having four major elements: content creation, content management, content access management, and production content delivery (IBM, 2002a). Content-wire.com (http://www.content-wire.com) includes 11 columns in a content life cycle: accessibility and usability, audiovisual, billing, content

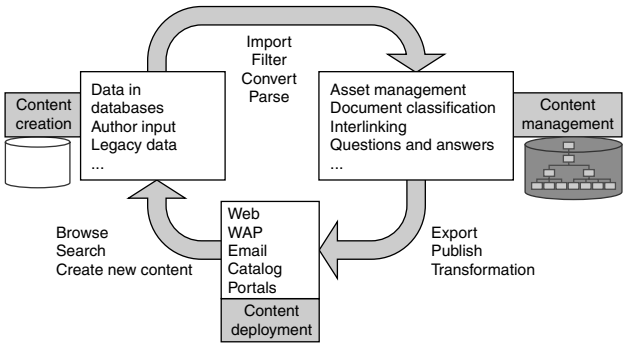


Figure 1: Web content life cycle.

delivery, content management, creation, digital copyright, syndication, taxonomies, unwired, and vertical content.

No matter how content management systems and practices vary, Web content must go through the processes of creation, management, and deployment (Figure 1). Content creation includes new content creation and existing content conversion such as data from databases and/or legacy content. Content management embraces many aspects such as

(1) asset management, i.e., managing all pictures and files to avoid data redundancies and easily handle media files;
(2) access management, which integrates user management and the permission systems;
(3) template management, which involves using dynamic page rendering mechanisms to allow a uniform presentation of content; and
(4) content representation and organization, which generates metadata and structured content.

Content deployment is the transfer of Web assets from the staging server to the production server. Current technology allows content on the production server to be broadcast not only to an organization's Web site, but also on other application devices. Because Web content is structured and organized in a way that can be reused, the system can render the same content into a variety of formats such as Web pages, wireless hand-held devices, e-mails, and catalogs, besides desktop browsers. Such content should also provide easier but more powerful browse and search capabilities for the front-end users.

CONTENT CREATION AND AUTHORING

Web content results from business processes such as planning, design, and implementation and falls into a number of large categories—depending on which criterion is used to categorize the content. For example, commonly seen content types include information on products, procedures and guidelines, reports from research or development projects or business transactions, presentations, and white papers to the public. While these content types serve different purposes in business processes, their creators may work at various positions (e.g., managers, programmers, graphic designers, technical writers) and have different levels of technology skills. Besides using good content authoring tools to leverage computer skill differences, it is also necessary to have predefined templates and procedures in place for content creators to follow. Content creation and authoring includes design, version control, collaborative authoring, and document management. These components work together to provide an infrastructure on which Web site functions and services are built.

Content Design

The general principles for content creation are to keep documents small, to build a modular system, to reuse content and definitions as much as possible, and to create powerful metadata (Gabriel, 2001). The size of display areas is limited on computer screens and smaller devices. Keeping documents small increases the displaying flexibility, but more importantly, smaller documents and document components increase the chance for them to be restructured and reused. It also makes it easier to frame the smaller documents into new technologies such as the extensible markup language (XML) (W3C, 2000). The second principle of modularization has been used successfully in building business and information systems. For large Web sites, modularizing means dividing Web content into modular systems based on certain criteria, e.g., by functional unit. A large site may have modules for research and development, customer service, and sales and marketing. Although different types of documents and content often bear different elements and structure, it is also quite common that some of them share the same definitions. An example would be news releases—each functional unit's Web site has a news section. Therefore, the document definition for news articles can be defined once and reused in all units' Web sites. Content and document definition reuse can save not only time and effort but also keep similar content in a consistent format. Metadata is information about data and documents. Title, author, key words, abstract, and content category are the most common elements for document type of content. For database content, metadata includes, among other things, table name, field name, and data type. Metadata contains elements underpinning the functions of browsing, searching, and displaying.

When designing Web content, one needs to consider both its internal and external structures. The internal structure refers to the organization and arrangement of content components. For example, an online technical manual may be structured as chapters, sections, and paragraphs; within each of the paragraphs there may be still or interactive illustrations with text annotation. An important design feature for such an online manual would be that no matter where users are inside the manual, there would be navigational indictors for them to move back and forth or up and down, even jump from one place to another. Another example is the navigation system within a Web site, which enables "relational" browsing and searching; i.e., similar products, technical support, and promotional plans pertaining to a particular product are related to that product's description. The role of internal structure design is to define Web content types and

structures and map out relationships among components of Web content.

The external structure of Web content refers to the way that Web content is presented to the front-end users. It defines which data components make up a Web page and how they will be arranged on the screen: dynamic posting, JavaScript-enabled interaction, or a drop-down menu. External structure design should also be compliant with the accessibility guidelines proposed by the World Wide Web Consortium (W3C). Nonetheless, these guidelines can also be applied to general external structure design. W3C recommends to

- Separate (internal) structure from presentation;
- Provide text so that text can be rendered in ways that are available to almost all browsing devices and accessible to almost all users;
- Create documents that work even if the user cannot see and/or hear; and
- Create documents that do not rely on one type of hardware (W3C, 1999a).

High-quality content requires planning and decisions on the sources, formats, and versions of content before it is created. Generally, Web content can be divided into four broad categories based on its source: internally created and owned, owned outright (e.g., commissioned artwork), acquired externally through leasing or subscription, or linked to external sources but not owned nor leased. The internally created and owned content includes a wide variety of documents as results of planning, analysis, designing, briefing, and many other organizational activities. These documents may be in the form of memos, reports, press releases, product catalogs, online transaction logs, white papers, or technical specifications and created as HTML or files in other formats (e.g., sound and images captured in audio/videos, or recorded as data sets). Some of them are currently active, i.e., being written or revised, while others may be inactive and eventually removed from the system. Some content such as planning and technical documents often involve collaborative creation and reviewing; hence the content may be privileged and need version and access control to ensure the efficiency, consistency, and confidentiality of the content under development. Both types of content—internally created and owned outright—are assets of an organization, though the organization does not have the right to revise or change the commissioned content. Content acquired externally usually comes in a package such as a database with its own user interfaces, whereas linked content can be anything and beyond the control of the Web content management system.

Version Control and Collaborative Authoring

The term "version control" has long existed before the Web. It was originally used to manage versions of source code in programming projects (Fogel, 2001). Version control software allows a developer or a group of developers to keep track of versions of program files that modifications have been made to. If a bug emerged after the new version was implemented, the developer can reinstate the earlier version into the system. If several developers are working on the same source code, the system can detect the changes and differences and automatically merge different versions into one copy. Version control in Web content creation and authoring essentially works the same way as a version control system. It records the history information about a file or directory, including things such as creation date, who created it, and the version number or label. The version control system allows authorized users to check out or check in files or directories. While a file or directory is checked out of the system, the user can lock it to prevent other users from accessing the same file or directory. After the updates are done, the user unlocks the file or directory by checking it back into the system. If a piece of content needs to be removed or added for any reason—or simply an earlier version of the Web site is preferred—the version control system can be used to restore the entire site to any previous state, rolling back multiple variations and edits by all authors until a satisfactory site can be put back in place (Dreilinger, 1999).

The concurrent versions system was implemented in a UNIX environment and intended for computer professionals. In the Web development environment, version control is as important as it is in software development. However, Web authors often are not computer professionals and come with various levels of computer skill. Besides, the content that Web authors create does not always fall into the category of program source code. The version control system must be compatible with various platforms and authoring tools. If a large team of Web developers works on the same Web project, "[t]he assets that compose a Web site must be factored in a way that allows many members of the Web team to make changes concurrently" (Nakano, 2002). The nature of Web content creation and authoring calls for user-friendlier version control and authoring tools.

Collaborative authoring is so tightly intertwined with concurrent version control that it is difficult to discuss one without mentioning the other. The concept of collaborative authoring encompasses a larger domain than version control. In addition to varied skill levels, Web content authors in a distributed environment often use different authoring tools on various platforms. When they work on the same project and need to share the same documents, the version control system needs to be interoperable besides accommodating different skill levels. To support collaborative authoring, the system needs to meet the requirements of equal support for all content types, concurrency control support, support for metadata, support for content-type independent links, retrieval of unprocessed source for editing, namespace manipulation, and support for collections (Whitehead, & Goland, 1999). In addressing these challenges, the Web Distributed Authoring and Versioning (WebDAV) working group of the Internet Engineering Task Force (IETF) developed the WebDAV protocol in support of the remote collaborative authoring on the Web (Goland, Whitehead, Faizi, Carter, & Jensen, 1999). Figure 2 demonstrates how authors in different locations using different authoring tools can edit the same set of documents with WebDAV server's version control functions.

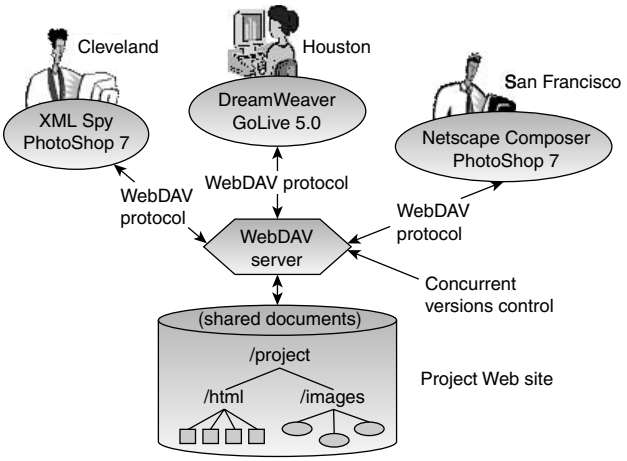


Figure 2: Collaborative authoring in three different locations and each Web author uses a different set of authoring tools.

Content authoring with concurrency control can make content creation efficient and ensure its consistency and quality. Authoring tools allow content creators to set templates and create consistent content structures for easy deployment. There is a wide variety of authoring tools available on the market, ranging from HTML to XML editors, many of which have capabilities of facilitating collaborative authoring.

Document Management Systems

Managing documents—e-mail messages, spreadsheets, image files, media files, or textual material—traditionally uses file systems in which documents are organized by location in hierarchies. The hierarchical structure of folders maps out the semantic structure of the file system. However, the hierarchical file system is not without a problem. For instance, documents can appear only in one place in the structure, even though they may play different roles and be relevant to other activities. As Dourish et al. (2000) described, the ideal document management system is the kind that provides a logical structure for document storage (meeting the needs of the system) while supporting document interaction (meeting the needs of users). Many commercial document management systems are built based on document properties, rather than document locations, because document properties are "the primary, uniform means for organizing, grouping, managing, controlling, and retrieving documents" (Dourish et al., 2000, p. 142).

A document management system may be a collection of roughly related systems that perform one or more of several functions, or a system consisting of a number of functional modules. Typical document management systems perform the following functions, though no one system performs them all:

- File naming: allowing long, descriptive file names;
- Indexing: assigning or extracting keywords and compiling them into lists;
- Multifile document control: presenting an electronic document consisting of multiple files in various formats to users as if it were a single entity;
- Storage and retrieval: managing the storage and retrieval of documents;
- "Library" services: performing functions of checking in, checking out, and versioning documents, auditing trails, and protecting documents from hacking, theft, and natural or human disasters;
- Workflow management: coordinating tasks, data, and content creators and developers to make the process more efficient and effective; and
- Presentation/distribution services: deciding the form and manner in which users view and interact with the content (Cleveland, 1997).

Increasingly, content developers use conceptual modeling methods to design the structure and relationships for a domain, in which data, documents, and applications are planned together in a systematic manner. These conceptual models sometimes are called "ontologies." Content-encoding schemas and application programs are then developed based on the conceptual models. Such schemas apply XML syntax and in some cases, the resource description framework (RDF) syntax (W3C, 1999b), which permits some inference or reasoning to allow intelligent agent applications.

CONTENT REPRESENTATION AND ORGANIZATION

Many Web sites today use representation and organization schemes of some sort. Yahoo! has its own category list that lets users browse its collection of Web sites hierarchically. Amazon.com uses a patented algorithm to recommend books or other goods based on the query a user entered for searching. Without a proper representation and organization of data and documents, it would have been impossible to provide these functions. Thus, the purpose of managing Web content is not simply to create some data and documents and make them available on the Web; more importantly, it is to create *organized* content so users can interact with it more effectively. Representation and organization needs to begin with content design and creation and follow through to the presentation (interface) on the Web. As Web content becomes increasingly complex and large in scope, several research communities such as library and information science and computer science have been actively studying how to represent and organize content to make it "machine-understandable" (W3C, 1999b). "Semantic Web" is a term describing the activities in this area, which includes a series of standards under development such as *Resource Description Framework (RDF) Schema Language* (W3C, 1999b) and *Requirements for a Web Ontology Language* (W3C, 2002a). Semantic Web is the abstract representation of data on the World Wide Web, based on the RDF standards and other standards to be defined. According to Berners-Lee et al. (2001), Semantic Web is "an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation." It is being developed by the W3C, in collaboration with a large number of researchers and industrial partners.

Content representation and organization refers to the process in which data and document models are created and relationships are mapped out as the architecture for templates and interfaces. Subsequently, such models are converted into logic designs and implemented into the system. Creation of representation and organization schemes and vocabularies requires methodologies from computer science, linguistics, library and information science, and other allied disciplines. There are two kinds of tools in content representation and organization: (1) metadata schemes that contain data elements describing various types of data sets and documents and (2) controlled vocabularies that are used to assign subject categories and indexing terms to documents.

Metadata Schemes

A metadata scheme defines a set of properties about data or documents. In the digital library community, there have been several proposals: the Dublin Core metadata sets (Weibel, Kunze, Logoze, & Wolf, 1998), the Warwick Framework (Lagoze, 1996), and the IEEE LOM/IMS metadata set for learning objects (IEEE/LTSC, 2002; IMS, 2002). The main purpose of metadata schemes is to provide a consistent way to record data about data sets or documents and to encode such data in a computer-understandable manner. The most common metadata elements include authors, titles, dates of creation/publication, owner/publisher, description, and category/subject terms. Some metadata schemes provide coarse-grained metadata elements to allow quick generation or creation of a metadata repository. Dublin Core is an example of this kind. With a structure containing three packages of content, intellectual property, and instantiation, it includes 15 elements and their subelements (Table 1). This set enumerates the very general yet essential data elements necessary for identifying and locating documents.

Metadata schemes may also include finer-grained data elements to describe documents in greater detail. For example, Dublin Core's educational extension, DC-ED, employs a set of elements designed specifically for describing the grade level, audience, pedagogy, and other details for a learning resource on the Web. In this sense, metadata schemes are the data models for collecting information about documents on the Web. Such information is used to search, browse, and locate documents on a topic over the Web. Another use of metadata schemes is to design them as templates to capture content directly from the Web. An example is Web-based forms, which have been widely used to capture customer data, survey data, sales orders, and best practices, to name a few.

Sometimes no standard metadata scheme is suitable for representing the types of data and documents being created. In such circumstances, a customized metadata scheme(s) is needed. The customized scheme may be created based on a standard such as Dublin Core, with a set of extended elements. It may also be created from scratch without using any existing standard. However, for interoperability, a metadata standard should be adopted whenever possible. Since metadata schemes are the data models for collecting information about data and documents, they often need to be embedded in content authoring tools so that content authors can create metadata while they are creating documents. In either case, i.e., creating an extended metadata scheme or a scheme for capturing data in a distributed environment, careful design is necessary to guarantee a sound representation of the data to allow easy retrieval and browsing of the content. Designing metadata schemes usually involves a data modeling process.

Modeling is a process of synthetic analysis and abstraction, in which system developers "construct an abstract description of a system in order to explain or predict certain system properties or phenomena" (Schreiber, Akkermans, Anjewierden, Hoog, Shadbolt, & Wielinga, 2000, p. 128). The result of this process is a model in which classes of objects in the real world are specified as having a transparent and one-to-one correspondence to an object in the model. Figure 3 is an example of the data model for an experience factory in the domain of workforce development.

The model in Figure 3 provides a high-level view of main classes and the relations among the classes. In the domain of workforce development, participating classes include the government who sponsors workforce programs such as "Welfare-to-work" and "Study-to-career," workforce organizations that initiate projects to execute programs, people in organizations who prepare documents describing and disseminating project information and results, and knowledge captured for sharing promising practices and lessons learned. A data model such as this serves as a communication tool for users of the Web content. It shows what kinds of content there will be on the Web site and how each class is related to one another. Before a data model reaches its acceptable version, discussions are often held with constituents to solicit feedback. As with any system development process, this discussion–revision process is iterative.

Table 1 Elements in Dublin Core

Content	Intellectual property	Instantiation
Title	Creator	Date
Subject	Publisher	Type
Description	Contributor	Format
Source	Rights	Identifier
Language		
Relation		
Coverage		

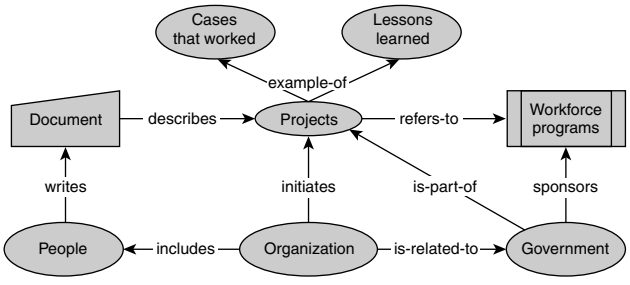


Figure 3: A sample conceptual model for the workforce services domain.

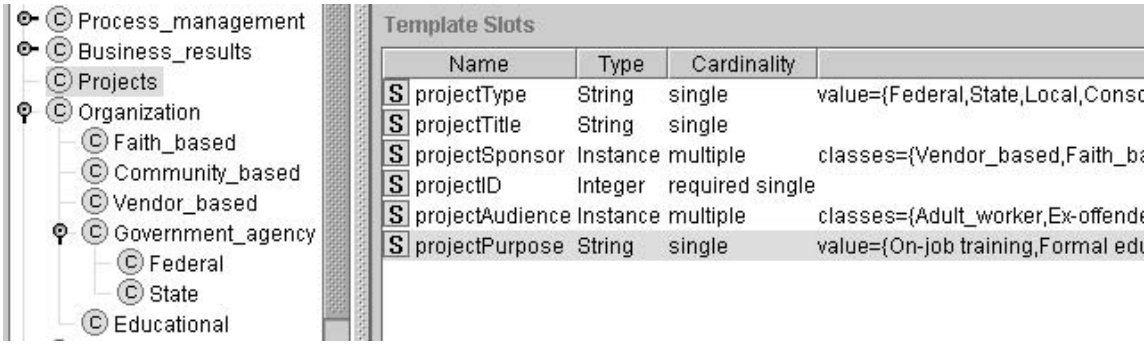


Figure 4: Portion of the workforce ontology.

Once the data model stabilizes, it needs to be translated into a structure with more details. Such a structure is sometimes called "knowledge structure" or "ontology" (not in the sense of philosophical studies of the world existing, but in the sense of representing the real world). Depending on the encoding language and software used, the same knowledge structure may appear differently. Figure 4 shows a portion of the ontology built from the data model in Figure 3. The *project* class in the workforce domain has a set of properties, each of which has name, type, constraints (cardinality), and other facets if the property type is symbol, class, or instance. For example, the property "project sponsor" uses "instance" as the property type. The class for property type "instance" is "Organization," from which instances will be extracted as the values for project sponsor. Organization as a class has its own set of properties. When it is used in another class (in this case, the project sponsor property in the project class), the class organization and its properties can be conveniently referenced and reused, thus saving time from redeveloping the same knowledge structure and improving consistency in representing the same class in different locations within the domain. The template properties provide a framework for building schemas that will be used to design interactive Web forms to capture data. Using the example in Figure 4, we can easily produce a database schema or an XML schema that can be used to create XML instance documents.

Controlled Vocabularies

Controlled vocabularies are tools for content classification and indexing. Taxonomies, thesauri, classification schemes, and glossaries are the forms of controlled vocabulary frequently used in categorizing and indexing content. Taxonomy is a term borrowed from biology. It implies three meanings according to WordNet (Princeton University Cognitive Science Lab, 1998):

(1) a classification of organisms into groups based on similarities of structure or origin;
(2) a study of the general principles of scientific classification; and
(3) the practice of classifying plants and animals according to their presumed natural relationships.

Taxonomies provide a method for structuring and classifying things—living organisms, products, subjects—into a series of hierarchical groups to make them easier to identify, study, or locate (Montague Institute, 2002). For digital content, taxonomies play two roles: as category labels for indexing content and as a tool for searching, browsing, and navigation. Category labels may be assigned manually by human content/metadata creators or automatically by using computer programs. The automatic approach often involves categorizing sample documents manually and then uses the manual result to train the system to classify other documents automatically. Properly classified content will allow an organization to inventory and monitor the Web content based on a structured understanding of user and community needs.

Taxonomies may be generated through automatically extracting, analyzing, and categorizing structured and unstructured content. The key question is how to extract different types of content and categorize them. Approaches commonly used in taxonomy development include statistical and symbolic methods (Parsons & Wand, 1997). The statistical (similarity-based) approach uses similarity measures to classify real world objects, which can be done automatically through computer programs. It is less labor intensive and time consuming compared to manual work, but also less likely to produce the organizational schemes the way people would. The symbolic (goal-oriented or explanation-based) approach relies on domain knowledge to determine the classes. Semantic networks (Woods, 1991) are examples of the symbolic approach. No matter which approach one uses, it is a good idea for taxonomy development to start with existing reference books and encyclopedias in the domain. The benefit is that using existing sources creates a familiar taxonomy that exhaustively covers a topic.

In the example as shown in Figures 3 and 4, the taxonomy for workforce services used a symbolic approach. By collecting concepts and terminology from workforce documents and reference sources, the initial taxonomy produced a three-level hierarchy of classes for the concept Project:.

```
Project
   Activities
      Case management
      Community audits
      Continuous improvement
      Cross training
      Employer focused programs
```

```
    Innovative research and studies
    Innovative technology
    Rapid response activities
    Regional strategic planning
    Rural solutions
    Staff collaboration
    Successful marketing strategies
    Training program
    Urban/Suburban solutions
    Youth development
  Service area
    City
    County
    Multi-cities
    Multi-counties
    Multi-states
    National
    State
    Tribal
```

Other controlled vocabularies such as thesauri and glossaries function similarly as taxonomies in slightly different ways. Thesauri are particularly useful in representing term relationships. Through the "broader term," "narrower term," and "related term" references, each term in a thesaurus is placed in a covert hierarchical structure and related to similar concepts. Glossaries, thesauri, and taxonomies are often used together to specify data definitions, preferred terms for indexing, and domain subject classes.

Encoding of Metadata and Taxonomies

Metadata schemes and taxonomies provide the semantics for representing data about Web content. As any digital content, they need a format for the computer to understand and manipulate. Conventionally, data elements in a metadata scheme are converted into data structures in a database. Taxonomies are also stored in a database format. The current trend is to encode these representations in XML format. XML describes a class of data objects called XML documents and partially describes the behavior of computer programs that process them (W3C, 2000). Unlike HTML that codes data for display (Listing 1), XML codes data for processing (Listing 2). Documents coded in XML format can tell the computer what content the data in between the tags is, so that the program can invoke the right processing instruction for this piece of data. Metadata coded by XML does not include any presentation information—it is separate from presentation. The main benefit of this separation is that it allows more semantics in the data representation and ease of data communication among systems. Another benefit is that the XML-coded data bears some structure through nesting (Listing 3). In Listing 3, the Book element may repeat to represent more books and the Author element to represent more authors if more than one author wrote the book.

```
<h1>Workforce development</h1>
<h2>Employer partners</h2>
<p>Quest Corporation</p>
<p>Riceland Foods, Inc.</p>
<h2>Program performance</h2>
<p>number of graduates: 230</p>
<p>Program duration: January 2000-present
  </p>
```

Listing 2: HTML Code for Data Display

```
<program>
   <title>Workforce development</title>
   <employerPartner id="1">Quest
     Corporation</employerPartner>
   <employerPartner id="2">Riceland Foods,
     Inc.</employerPartner>
   <performance>
      <numberOfGraduates>230
        </numberOfGraduates>
   </performance>
   <programDuration>January 2000-present
     </programDuration>
</program>
```

Listing 3: XML Code for Data Processing

```
<Catalog>
     <Book>
         <Title>XML Stylesheet Language
           </Title>
         <Authors>
            <Author>John Brown</Author>
         </Authors>
         <Publisher>AZ Publishing
           </Publisher>
         <PubDate>May 2002</PubDate>
         <ISBN>1-234567-89-0</ISBN>
     </Book>
</Catalog>
```

Listing 4: XML Document for the Book Catalog Example

CONTENT TRANSFORMATION

The ultimate goal of representing and organizing Web content is to deliver the content effectively and efficiently, so that users can retrieve the right content at the right time. Web content can be delivered based on personal profiling, the task to be performed, or the information to be communicated. The delivery is a technology-intensive process since documents and data stored in databases or XML file systems need to be rendered into the formats viewable via a desktop browser, mobile phone, automobile-based personal computer, or hand-held PC. The content is also expected to be able to serve multiple purposes across enterprise Web sites. For instance, the product data may be rendered into an online catalog, shown in an online purchase order, or delivered as a marketing e-mail.

XML documents can be presented through two kinds of stylesheets: the cascading style sheet (CSS) and the extensible stylesheet language (XSL). CSS is divided into Level 1 and Level 2, both of which are W3C recommendations. CSS can be used as internal stylesheets embedded in HTML documents or external URIs linked to the

stylesheet documents. When CSS is used to define formatting styles for a tag or class of tags, it automatically applies the styling and overrides any default styling that may apply (such as the formatting tags H1, H2, etc.). Although CSS is useful for formatting and determining how a document will look, it is a strict language in the sense that the source content can only appear in the order as it is. In other words, no reordering is allowed and hence offers nothing for structural transformation.

XSL is a language that provides the mechanism to transform and manipulate XML data. The XSL specifications consist of a number of components: XSL transformation (XSLT), XSL formatting objects (XSL-FO), and XML path language (XPath) (W3C, 2002b). The benefits of using separate styles to transform content include

- Reuse of fragments of data: the same content should look different in different contexts;
- Multiple output formats: different media (paper, online), different sizes (manuals, reports), and different classes of output devices (workstations, hand-held devices);
- Styles tailored to the reader's preference (e.g., accessibility): print size, color, simplified layout for audio readers;
- Standardized styles: corporate stylesheets can be applied to the content at any time; and
- Freedom from style issues for content authors: technical writers needn't be concerned with layout issues because the correct style can be applied later (Grosso & Walsh, 2000).

Unlike CSS, which is limited in presenting XML data, XSLT describes rules for transforming a source tree (as defined in the form of a document-type definition (DTD) or an XML schema) into a result tree. The result tree may or may not bear the same structure as the source tree because XSLT specifications permit transformation templates to reorder the sequence of XML data elements, create new elements, and manipulate loop processing of XML data. By associating XML document structures with XSL templates, the transformation matches the structural patterns against elements in the source tree and generates a result tree that is separate from the source tree (Figure 5) (W3C, 1999c). A transformation expressed in XSLT is called a stylesheet. This is because, in the case when XSLT is transforming into the XSL formatting vocabulary, the transformation functions as a stylesheet. Listing 4 shows the template for transforming the XML instance into an HTML document.

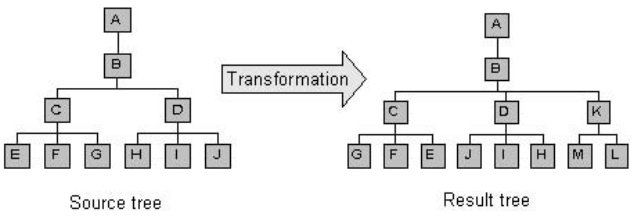


Figure 5: An example of XSL transformation: the data elements are reordered and new elements are created in the result tree.

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE xsl:stylesheet SYSTEM "C:\box5.
  dtd">
<xsl:stylesheet version="1.0" xmlns:xsl=
 "http://www.w3.org/1999/XSL/Transform"
xmlns fo="http://www.w3.org/1999/XSL/
  Format">

   <xsl:template match="/">
       <html>
          <head>
             <title>Books in Stock</title>
          </head>
          </body>
             </xsl:apply-templates/>
          </body>
       </html>
   <xsl:template>

   <xsl:template match="Book">
       <h2>
          <xsl:apply-templates select=
            "Title">
       </h2>
   </xsl:template>

   <xsl:template match="Authors">
       <p>
          <xsl:for each select="Author">
             <xsl:value-of select=
               "Author"/>
          </xsl:for-each
          </xsl:apply-templtes/>
       </p>
   </xsl:template>

   <xsl:template match="Book">
       <xsl:value-of select="Publisher"/>
       <xsl:value-of select="PubDate"/>
       <xsl:value-of select="ISBN"/>
       <xsl:apply-templates/>

   </xsl:template>

</xsl:stylesheet>
```

Listing 5: The Transformation Stylesheet for the Example in Listing 4

In Listing 4, the first template creates an HTML shell for the content and the second and third template rules transform the title, author, publisher, publication date, and ISBN into HTML elements. Compared to CSS, XSLT offers much more capabilities in manipulating XML data.

To complete the transformation from XML to various deliverable presentations, XSLT must work together with two other component languages—XPath and XSL formatting language. XPath is used to reference specific parts of an XML document while the XSL formatting language describes a set of formatting objects and properties.

ISSUES IN CONTENT MANAGEMENT

As Web sites play an increasingly important role in organizational activities and business processes, major issues in Web content management need to be dealt with early in the planning process. Such issues include content and technical requirements, access control, economic and legal aspects, and security.

Content Requirements

It is essential that a Web site maintain quality and timely content. This means that a content management system needs to

(1) provide fresh, up-to-date, accurate, and personalized content;
(2) allow efficient authoring for nontechnical content providers;
(3) support worldwide content authoring in different languages and at different locations;
(4) provide automated scheduling for both content publishing and archiving;
(5) integrate workflow processes to automate content approval;
(6) build a component architecture separating content from presentation forms and dynamic serving of content; and
(7) support version archiving and an audit trail to provide a record of site changes (McCluskey, 2000).

Decisions must be made in dealing with time-sensitive content, e.g., how often content should be updated—whether to update on a daily or hourly basis and for what type of content. Different types of content also vary in terms of their stability. For relatively stable content, e.g., reports and procedures, the content structure may adopt some stable and consistent document structure and style. The workflow from creating the content to delivering it to the Web site may need to go through several approval steps: when an update or new content is to be posted, how many approvals should it go through? Can content be updated in real time? Will certain content be updated on a scheduled basis? All these issues need to be solved before a content management system is built or customized.

Technical Requirements

Technical support for a content management system includes the Web management team's collective technical strength and capabilities, platforms for system and networking environment, and software applications. To operate a complex Web site and keep the content up-to-date, the availability of technical resources (Web designers and developers) is extremely important. Issues such as authoring environment (Windows, Macintosh, Solaris), content repositories (SQL database server, Oracle, Sybase, etc.), and authoring tools (Netscape Composer, Dreamweaver, etc.) should also be considered.

Access Control

An organization's Web site generally contains content with three modes of access: open-access information, license-restricted information, and privileged information (Sullivan, 2001). Open-access information is freely available to all users. News and press releases, product catalogs, technical support, and other publicly available information fall into this category. Access to license-restricted information restricts the retrieval and use of such sources to the members of an organization(s), such as commercial databases that an organization subscribed to, or the entire digital library of a professional association. In these cases, access control can be done through basic user authentication or server address verification.

The most challenging in access control is to the privileged content, which can range from data about individual customers collected through Web forms to business intelligence about a company. The access to this type of content is controlled by the need to know. Examples include granting access to a project's information to its team members and the project leader may require access to other projects related to this one.

Economic and Legal Aspects

When an enterprise produces information, a high cost is associated with the information production process. The economic goal of Web content management is to pay for the least cost to gain the most efficiency in creating, organizing, and delivering the content. The return-on-investment for Web content management may not be easily translated into monetary terms; other nonmonetary measures such as time saved, decision quality improvement, and reduced effort can be useful measures.

Legal aspects related to Web content management mainly focus on public versus nonpublic content as well as digital rights management. For the U.S. Government, all data collected, created, received, maintained, or disseminated by any of its agencies must be made accessible to the public unless the data are classified as inaccessible by state or federal statute. Written policies and procedures are also required to assure properly controlled access to private and confidential data. Enterprises need to be compliant with the access and security laws when it comes to private and confidential content. While the philosophy for enterprise Web content management may be different from that of the government, the privacy and confidentiality rules apply to all organizations.

The term "digital rights management" (DRM) emerged in the past couple of years to cover "the description, identification, trading, protection, monitoring and tracking of all forms of rights usages over both tangible and intangible assets including management of rights holders relationships" (Iannella, 2001). For digital content, DRM controls not only who has access to certain content, but also determines the amount of access, e.g., for a limited time, a limited number of times, or a certain number of accessing machines or users. It also manages payment for uses of digital content. This area is gaining more attention in the information industry.

Security

It is common for enterprise Web sites to use firewalls, intrusion detection software, or content filters to protect

them from outside attacks (e.g., hackers, viruses) and to prevent cyberthieves from stealing intellectual property within corporate IT systems. However, these firewalls offer only a limited line of defense because data and communication must flow between the inside and outside of the security line. As ports open for e-mails, Web, and other types of content to pass through, holes are left open for security breaches. Security for Web content management is an integral part of the enterprise IT system. It includes four major areas: identity management, access management, threat management, and privacy management (IBM, 2002). Identity management deals with user ID, password, and account setup and change, while access management needs to support any type of user authentication and to control access to any type of resource from the authenticated users. When the system is intruded upon or attacked, the threat management function should be able to quickly determine the severity of attacks and issue alerts to the whole system. Some of the security areas (e.g., privacy management) overlap with access control and legal aspects of content management mentioned earlier.

TRENDS

Web sites at present are becoming increasingly complex. Managing complex interrelationships among these sites and documents inside them requires a collaborative effort between developers, content contributors, and Web professionals. The Gartner Group predicted four key trends in Web content management:

(1) it is becoming a core management tool supporting e-business and other critical applications;
(2) vendors will begin to consolidate disparate technologies into their offerings, such as rich media support, catalog management, syndication, and personalization;
(3) the Web content management market is filled with small or financially troubled vendors that will merge, be acquired, or simply cease operations in the coming year; and
(4) prices have come down dramatically, and will continue to fall down (Latham, 2002).

While this forecast focuses more on the marketplace of Web content management, it also implies some of the more specialized areas developing rapidly.

One of such specialized areas is digital rights management (DRM). By definition, DRM represents the concepts, systems, and functions that enable content owners to distribute various forms of digital assets securely, maintain visibility to its creators, and determine the means by which that content can be used, reused, purchased, copied, and distributed by its users (Wood, 2002). The content value chains make DRM increasingly important in a content management system. Open source is another area gaining popularity on the marketplace for content management systems. Open source usually includes packages and tools distributed free of charge under a license that guarantees the right for developers to read, redistribute, modify, and use the software (source code and all) freely. Open-source architectures are popular not only because they are free, but also because they tend to be fairly stable and some major open-source projects provide access to a vast network of global developers, who offer support and share knowledge around the clock.

Knowledge-driven organization of Web content has been a hot topic in the past few years. As XML, RDF, and ontologies are being developed and adopted by more and more Web sites, Web developers and content authors are expected to create and deliver the content using templates built based on knowledge models. Knowledge structures such as ontologies and taxonomies have become the underpinning infrastructure for the content, structure, and presentation of a Web content management system.

GLOSSARY

Collaborative authoring A field of computer-supported collaborative work (CSCW) concerned with tools that allow teams of two or more to create a document together: examples include a 10-page paper, a blueprint of an industrial plant, a million-line computer program, and a hypertext encyclopedia with hundreds of thousands of entries (Vitali, 2002); also known by the terms computer-supported collaborative writing, cooperative writing, cooperative editing, shared editing, and group editing.

Concurrent versions system (CVS) The dominant open-source network-transparent version control system; useful for everyone from individual developers to large, distributed teams. It uses the client–server access method and allows developers to access the latest code from anywhere there's an Internet connection. Its features, such as the unreserved checkout model and version control, avoid artificial conflicts common with the exclusive checkout model (CollabNet, 2002).

Content transformation The transformation of an XML document into HTML, XHTML, or other forms that can be presented through a viewing device by writing an XML stylesheet language (XSL) template and formatting rules in order to deliver the document's content to users while maintaining its structure.

Digital rights management (DRM) The concepts, systems, and functions that enable content owners to securely distribute various forms of digital assets, maintain visibility to its creators, and determine the means by which that content can be used, reused, purchased, copied, and distributed by its users.

Document management The process of managing documents through their life cycle from inception through creation, review, storage and dissemination all the way to their destruction (Document Management Avenue, 2002).

Staging Web site A copy of the production Web site that developers use to test before the changes or new content goes live on the production site.

Taxonomy A method for structuring and classifying things—products, customers, and services—into a series of hierarchical groups to make them easier to identify, study, or locate.

Version control The systematic maintenance and organization of all the versions of a set of files made over time by allowing people to consult previous revisions of individual files to compare and view the changes made between them, thereby keeping an accurate and retrievable log of a file's history and, more importantly, enabling people (even if in geographically disparate locations) to work together on a development project over the Internet or private network by merging their changes into the same source repository (Helixcommunity.org, 2002).

Web content life cycle The processes of content creation, representation and organization, transformation, and delivery.

Web content management (WCM) A process by which WCM solutions are used to create, store, publish, update, and repurpose content to be communicated through an organization's Web site by way of templates, workflow tracking features, publishing systems, and storage of content objects in database or file indexing systems. Web content may appear in the form of an HTML, PDF, or Word document, PowerPoint slide decks, Flash, XML, animation, interactive/still image, audio/video, plain text, database, etc.

Web content representation and organization Two aspects of the content management process: the modeling of data or document components and their properties, which will be used to create templates and schemas for content creation; and the modeling of the domain subject in the form of classification or taxonomy, which will be used to categorize data or document components for content retrieval and browsing.

WebDAV (Web-based distributed authoring and versioning) A set of extensions to the HTTP protocol that allows users to collaboratively edit and manage files on remote Web servers (Internet Society, 1999).

Workflow A combination of people, projects, and business environment in which Web content tasks are carried out from the initial content producers to designers, editors, and reviewers along the content production line until such tasks are completed.

XML (extensible markup language) A class of data objects called XML documents; also partially describes the behavior of computer programs that process them (W3C, 2000).

XML stylesheet language transformations (XSLT) A language that provides the mechanism to transform and manipulate XML data (W3C, 1999c).

CROSS REFERENCES

See *Cyberlaw: The Major Areas, Development, and Provisions; Extensible Markup Language (XML); Extensible Stylesheet Language (XSL); Web Site Design.*

REFERENCES

Berners-Lee, T., Hendler, J., & Lassila, O. (2001, May). The semantic Web: A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities. *Scientific American*. Retrieved November 30, 2002, from http://www.scientificamerican.com/article.cfm?articleID=00048144–10D2–1C70–84A9809EC588EF21&catID=2

Cleveland, G. (1997). Document management systems. Retrieved November 30, 2002, from http://www.nlc-bnc.ca/9/1/p1-243-e.html

CollabNet. (2002a). Concurrent versions system: The open standard for version control. Retrieved November 30, 2002, from http://www.cvshome.org/

Document Management Avenue (2002). Document management glossary. Retrieved November 30, 2002, from http://cgi.parapadakis.plus.com/modules.php?name=Encyclopedia&op=list_content&eid=1

Dourish, P., Edwards, W. K., LaMarca, A., Lamping, J., Petersen, K., Salisbury, M., et al. (2000). Extending document management systems with user-specific active properties. *ACM Transactions on Information Systems, 18*(2), 140–170.

Dreilinger, S. (1999). CVS version control for Web projects. Retrieved November 30, 2002, from http://www.durak.org/cvswebsites/howto-cvs/index.html

Fogel, K. (2001). *Open source development with CVS* (2nd ed.). Scottsdale, AZ: Coriolis Group Books.

Gabriel, J. (2001). Content management: designing a robust XML-based content management system. *XML Journal, 2*(7), 48–52.

Goland, Y. Y., Whitehead, E. J. Jr., Faizi, A., Carter, S., & Jensen, D. (1999). HTTP extensions for distributed authoring—WebDAV. Retrieved November 30, 2002, from http://www.webdav.org/specs/

Grosso, P., & Walsh, N. (2000). XSL concepts and practical use. Retrieved November 30, 2002, from http://www.nwalsh.com/docs/tutorials/xsl/xsl/frames.html

Helixcommunity.org (2002). A version control glossary. Retrieved November 30, 2002, from https://www.helixcommunity.org/nonav/docs/ddCVS_cvsglossary.html

Iannella, R. (2001). Digital rights management (DRM) architecture. *D-Lib Magazine, 7*(6). Retrieved from http://www.dlib.org/dlib/june01/iannella/06iannella.html

IBM (2002). Content management. Retrieved November 30, 2002, from http://www-1.ibm.com/services/kcm/cm_wcm.html

IBM (2002). Establishing information security as a business enabler: How IBM helps deliver a secure infrastructure for e-business. Retrieved November 30, 2002, from http://www-3.ibm.com/software/tivoli/info/security/wp-security-enabler/index.jsp

IEEE/LTSC (2002). Draft standard for learning object metadata. Retrieved November 30, 2002, from http://ltsc.ieee.org/doc/wg12/LOM_1484_12_1_v1_Final_Draft.pdf

IMS (2002). IMS Learning resource metadata specifications. Retrieved November 30, 2002, from http://www.imsproject.org/metadata/

Internet Society. Network Working Group (1999, February). HTTP extensions for distributed authoring—WebDAV. Retrieved November 30, 2002, from http://asg.web.cmu.edu/rfc/rfc2518.html

Lagoze, C. (1996, July–August). The Warwick Framework: A container architecture for diverse sets of metadata. *D-Lib Magazine*. Retrieved November 30, 2002, from http://www.dlib.org/dlib/july96/lagoze/07lagoze.html

Latham, L. (2002). Web content management: content is your currency. Retrieved November 30, 2002, from http://www.tagtech.org/usr_doc/web%20content%20management.pdf

McCluskey, N. (2000). Untangling Web content management: Intranet, Extranet and otherwise. *Intranet Journal*. Retrieved November 30, 2002, from http://www.intranetjournal.com/text/articles/200004/im_04_18_00a.html

Montague Institute (2002). Indexes and thesaurus. Retrieved November 30, 2002, from http://www.montague.com:8080/Public/indexes.htm

Nakano, R. (2002). *Web content management: A collaborative approach*. Boston, MA: Addison-Wesley.

Parsons, J., & Wand, Y. (1997). Choosing classes in conceptual modeling. *Communications of the ACM, 40*(6), 63–69.

Princeton University Cognitive Science Lab (1998). WordNet: A lexical database for the English language. Retrieved November 30, 2002, from http://www.cogsci.princeton.edu/~wn/

Schreiber, G., Akkermans, H., Anjewierden, A., Hoog, R. de, Shadbolt, N. R., & Wielinga, B. (2000). *Knowledge engineering and management: The CommonKADS methodology*. Cambridge, MA: MIT Press.

Sullivan, D. (2001, August 31). Five principles of intelligent content management. *Intelligent Enterprise*. Retrieved November 30, 2002, from http://www.intelligententerprise.com/010831/413feat1_1.shtml

Vitali, F. (2002). Collaborative authoring on the Web. Retrieved November 30, 2002, from http://www.ktic.com/topic6/12_COLLA.HTM

W3C (1999a). Web content accessibility guidelines 1.0. Retrieved November 30, 2002, from http://www.w3c.org/TR/WAI-WEBCONTENT/

W3C (1999b). Resource description framework (RDF). Retrieved November 30, 2002, from http://www.w3.org/RDF/ (Date of access: November 30, 2002).

W3C (1999c). XSL transformations (XSLT) version 1.0. Retrieved November 30, 2002, from http://www.w3.org/TR/xslt

W3C (2000). Extensible markup language (XML) 1.0 (second edition). Retrieved November 30, 2002, from http://www.w3.org/TR/2000/REC-xml-20001006

W3C (2002a). Requirements for a Web Ontology Language: W3C working draft 08 July 2002. Retrieved November 30, 2002, from http://www.w3.org/TR/webont-req/

W3C (2002b). The extensible stylesheet language (XSL). Retrieved November 30, 2002, from http://www.w3.org/Style/XSL/

Weibel, S., Kunze, J., Logoze, C., & Wolf, M. (1998). Dublin Core metadata for resource discovery. Retrieved November 30, 2002, from ftp://ftp.isi.edu/in-notes/rfc2413.txt

Whitehead, E. J. Jr., & Goland, Y. Y. (1999). WebDAV: a network protocol for remote collaborative authoring on the Web. In *Proceedings of the Sixth European Conference on Computer Supported Cooperative Work* (pp. 291–310). Dordrecht, the Netherlands: Kluwer Academic Publishers.

Wood, J. (2002). Digital rights management (DRM). Retrieved November 30, 2002, from http://www.cmswatch.com/ContentManagement/Topics/FeaturedTopic/?topic_id=10

Woods, W. (1991). Understanding subsumption and taxonomy: a framework for progress. In J. Sowa (Ed.), *Principles of semantic networks: Explorations in the representation of knowledge* (pp. 45–94). San Mateo, CA: Morgan Kaufman.

Web Hosting

Doug Kaye, *RDS Strategies LLC*

INTRODUCTION

Web-hosting services (like Web sites) come in all shapes and sizes. The chart in Figure 1 shows the distribution of annual Web-hosting budgets for U.S. businesses. The average budget is on the order of $1,200 to $1,800 per year ($100 to $150 per month), but note that more than 10% of all businesses surveyed spend more than $100,000 per year.

Because of this tremendous range of offerings, you might think that the services at one end of the spectrum are very different from those at the other end. In fact, these services are far more alike than they are different. For example, all Web sites, no matter how small or how large, require Web servers, domain name services, backup and recovery, and connections to the Internet. But it would be nearly impossible to analyze as a single group this wide range of offerings that fall under the Web-hosting umbrella. In order to keep our analysis more manageable, we will segregate the vendors into categories, and then study each category in detail.

THE CATEGORIES

After a few years of confusion over the various types of Web hosting available, the vendors have settled into four distinct service categories. Nearly everyone in the Web-hosting industry and the trade press has accepted these categories. As a result, the categories are now consistent and helpful in distinguishing the many vendors.

In the least expensive, or low-end, category are *shared servers*. As their name implies, these are computer systems that are shared by more than one Web site, and hence are appropriate for small, simple, low-traffic sites.

Next on the list are *dedicated servers*. These are nearly identical to shared servers with the obvious exception that they are not shared but rather dedicated to a single Web site or to multiple Web sites owned and controlled by the same business entity. As compared to shared servers, dedicated servers offer more capacity and flexibility and better security, but at a higher price.

The next category is a substantial step away from the previous categories, but in some ways a step backward. Instead of offering more support than is available from shared- and dedicated-server Web-hosting services, *colocation* is a rather bare-bones service that merely houses servers in a data center and connects those servers to the Internet. It does not include the server hardware or any of the software and services necessary to operate a Web site. Colocation by itself is aimed at customers who want to supply and manage their own Web-site hardware and software, but who do not want to provide the physical facilities and may not want to manage their links to the Internet.

Managed service providers (MSPs)—the fourth and final category—address the huge gap between the bare-bones offerings of colocation services and the needs of owners of major Web sites. Colocation services and MSPs have developed truly symbiotic relationships in which one could not succeed without the other, and the combination of these two services is often the choice of high-end Web sites.

THE COMPONENTS OF WEB HOSTING

The *service component pyramid* in Figure 2 illustrates the relationships of the separate components of Web hosting. Each layer generally supports the layers above it and depends on the layers below it. For example, operating-system software provides an environment for application servers yet requires hardware on which to run.

Note the following:

The service components available from shared- and dedicated-server Web-hosting services are essentially the same and have, therefore, been combined into a single group.

Colocation, on the other hand, includes very few service components—just those at the lowest levels of the

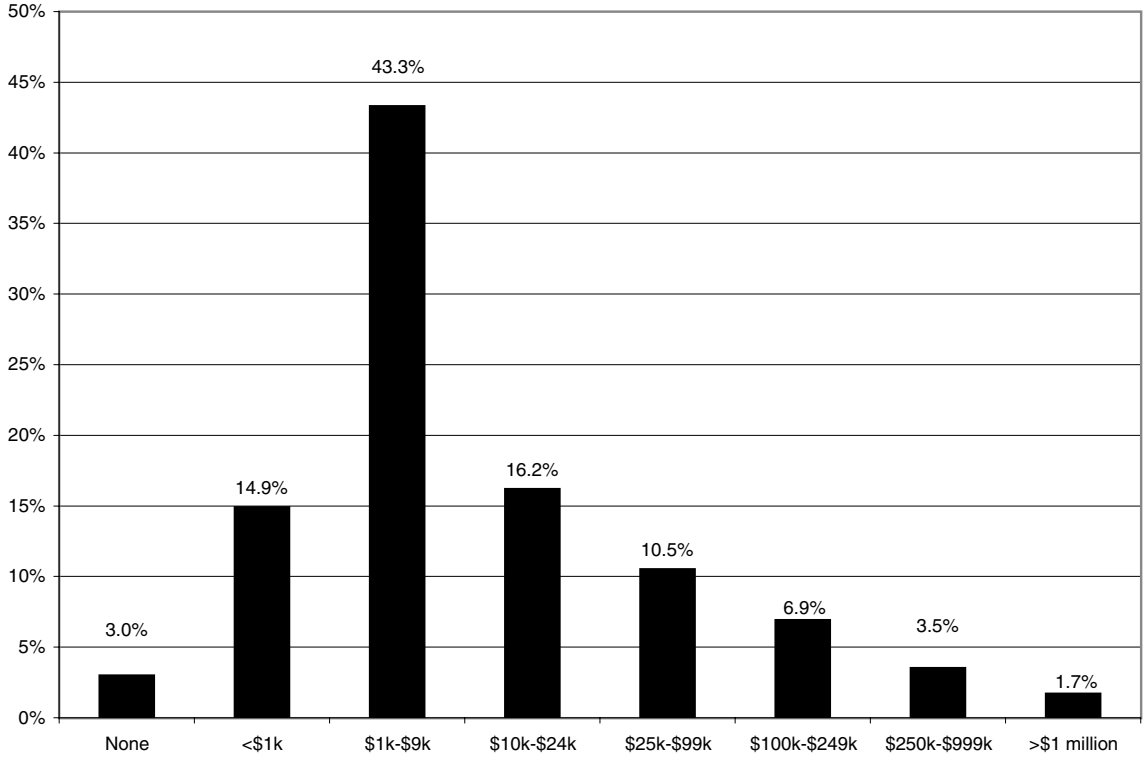


Figure 1: Annual Web-hosting budgets. Source: ActivMedia Research (www.activmediaresearch.com).

pyramid. When colocation is combined with managed services, however, the service components actually exceed those available with shared and dedicated servers.

The incremental difference between the shared/dedicated and colocation/managed service provider (MSP) combinations is the MSPs' support of application servers and databases. This is because the Web sites outsourced to MSPs are typically the largest and most complex. They are often based upon application servers that, in turn, depend on high-end database packages.

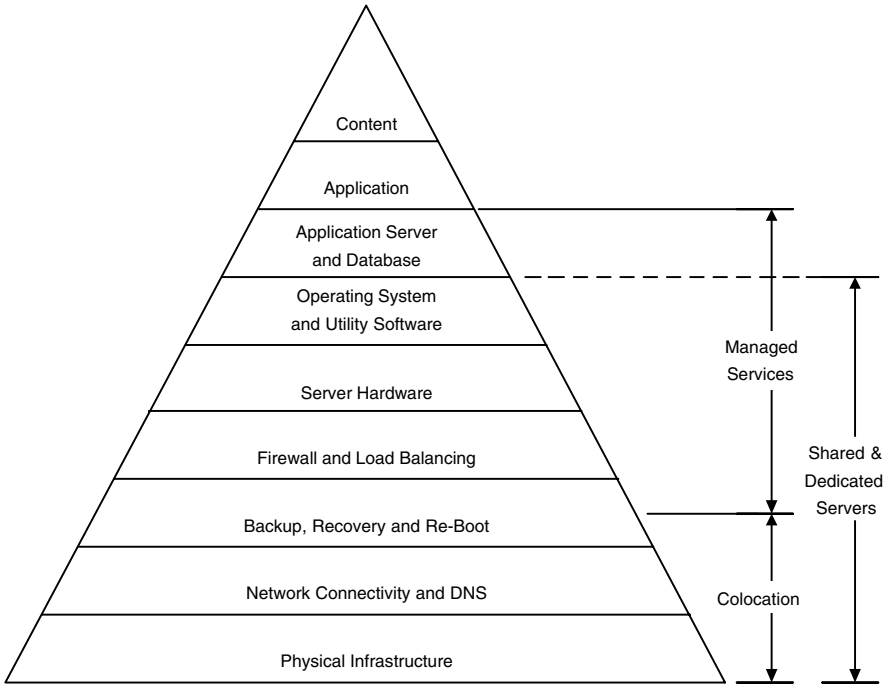


Figure 2: The service component pyramid.

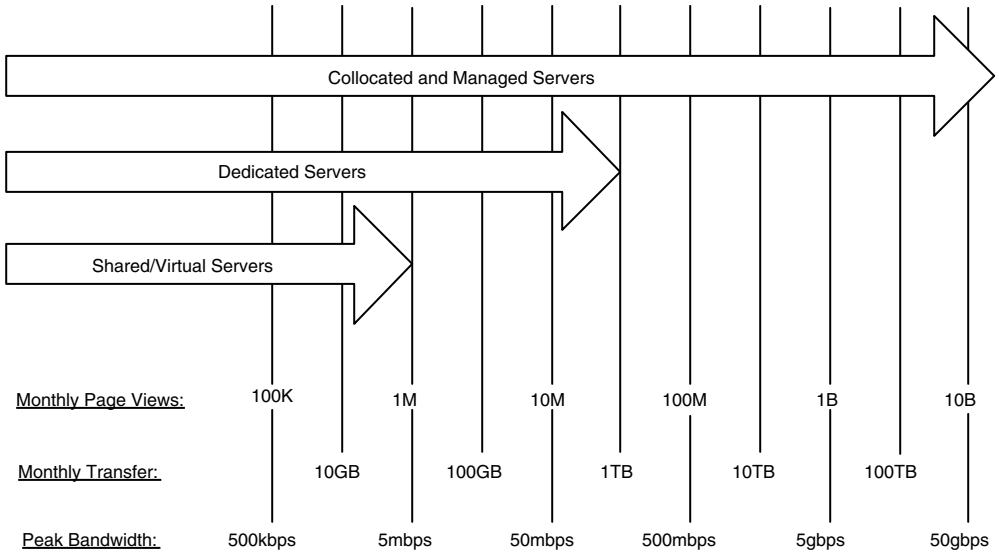


Figure 3: Service categories compared by traffic volume.

WEB SITE TRAFFIC

The second criterion used to segregate the four classes of Web-hosting services is the amount of Internet traffic their customers' Web sites generate.

Figure 3 allows one to pinpoint a Web site according to its traffic volume. For instance, if a site will deliver approximately 5 million page views per month, shared servers will be unsuitable. Dedicated servers or colocation (with or without managed services) should be considered instead. If a site will deliver only 100,000 page views per month, shared servers will likely be the best alternative.

All three of the traffic measurements (page views, transfer, and peak bandwidth) are related to one another (see Measuring Web Site Traffic and Bandwidth), but the relationships are not simple. If the monthly page views increase, for example, so will the volume of data transfer and the demand for peak bandwidth. But the ratios of these measurements (or *metrics*) to one another are not constant. For instance, two Web sites may each deliver one million page views per month, but if the pages on one site are twice the size of the pages on the other, the former site's data transfer will be twice that of the latter site.

MEASURING WEB SITE TRAFFIC AND BANDWIDTH

The measurements shown in Figure 3 are the three most commonly used in analyzing Web-site traffic volume.

Page views. The number of Web pages delivered to users. This metric is used most in Internet advertising, because page views or *impressions* of banner ads are one way by which ad campaigns are planned and measured. The most popular sites, such as Yahoo!, deliver on the order of 5 billion (5 B) page views each month.

Monthly transfer. The metric most commonly used by shared- and dedicated-server Web-hosting services. This is a count of all the data (measured in bytes) that a Web site's servers transfer to or from the Internet. If a user knows the average size of its Web site's pages, it can directly compute the monthly transfer as the number of page views multiplied by the average page size. For example, if a site's average page size is 40 K bytes, and the site delivers 1 million pages each month, it will transfer 40 billion bytes (40 gigabytes or 40 GB) per month. (Note that the average page size may need to be adjusted according to the frequency of delivery of pages, because some sites deliver large pages more frequently than small pages, or vice versa.)

Peak bandwidth. Used by colocation and managed service providers. If a Web site's connection to the Internet is envisioned as a pipe, *bandwidth* (the capacity of the pipe) is a measurement of the diameter of the pipe that is required to handle a site's traffic, whereas page views and monthly transfer are measurements of the actual number of data that flow through the pipe.

Think of the electrical service that enters a home or office: Opening the service panel might reveal a circuit breaker labeled "200 amperes." This is the maximum amount of electricity that can be used at one time and is like bandwidth. The bill each month from the utility company for the amount of electricity actually used is comparable to monthly transfer.

Just because a colocation service or MSP charges by bandwidth as opposed to transfer does not mean using this method of measurement will cost more. In fact, because sites hosted at colocation facilities or by MSPs tend to be larger and busier than those hosted by shared- or dedicated-sever vendors, their actual cost of connectivity to the Internet is typically less due to volume discounting.

Most colocation vendors and MSPs use the *95th percentile rule*. The most important factor to understand is that under the 95th percentile rule bandwidth is not billed according to the absolute peak, or *burst*, but rather according to the greatest usage that is sustained over certain units of time. Using the electricity comparison again, the 200-ampere circuit breaker will not trip just because of

the very short spikes that occur when, say, an air conditioner is turned on. Rather, the current must exceed 200 amperes for a certain period of time.

The computation of peak bandwidth from monthly page views is not as exact as the computation of monthly transfer because one Web site may deliver its page views evenly over a 24-hour period, whereas another may tend to be visited mostly during a narrower period of time during the day. An approximation of 1 Mbps (one megabit per second) of bandwidth for 200,000 page views per month is a good rule of thumb.

COST

We have looked at Web-hosting service categories from two perspectives: the traffic volumes they can handle and the service components they offer. The final comparison we will make is according to their costs, as illustrated in Figure 4.

The *x*-axis of Figure 4 indicates the total cost per month for hardware and software of an operational Web site. Note that the scale is logarithmic. The costs range from less than $10 to over $1 million per month.

In order to make an apples-to-apples comparison, Figure 4 includes the monthly cost of hardware and software it takes to run a Web site. Shared- and dedicated-server vendors (and in some cases, MSPs) own the server hardware and most of the software licenses. They rent or lease those components to their customers—hence the inclusion of these costs in the comparison. In the case of pure colocation, however (i.e., without the addition of third-party managed services), the customer must either own the hardware and software or lease these components through third parties. For colocation, therefore, the budgets shown include either the monthly leasing costs of these components or the equivalent depreciation or amortization.

The *y*-axis scale has no values or calibration points. Instead it is a relative scale of features and services. Being higher on the scale implies more features or a greater level of service.

We have now looked at the four categories of Web-hosting vendors from three perspectives: We have seen how they vary according to the services they offer, by the amount of Web site traffic they are each intended to handle, and by cost. Next, we will examine each category in detail.

SHARED AND DEDICATED SERVERS

A *shared server* (also referred to as a virtual server) is a single computer system on which a Web-hosting service runs multiple small Web sites owned by separate customers. The software for each Web site runs in a *virtual operating environment* that protects it from other Web sites running on the same physical server, and vice versa.

Shared servers are used for the vast majority of all Web sites. Because these Web sites require relatively low levels of computer resources, multiple sites—sometimes thousands—can be run on a single server. This means, in turn, that a shared-server hosting provider can offer Web hosting for as little as $19.95 per month and still make money. Low-cost shared-server hosting can be an

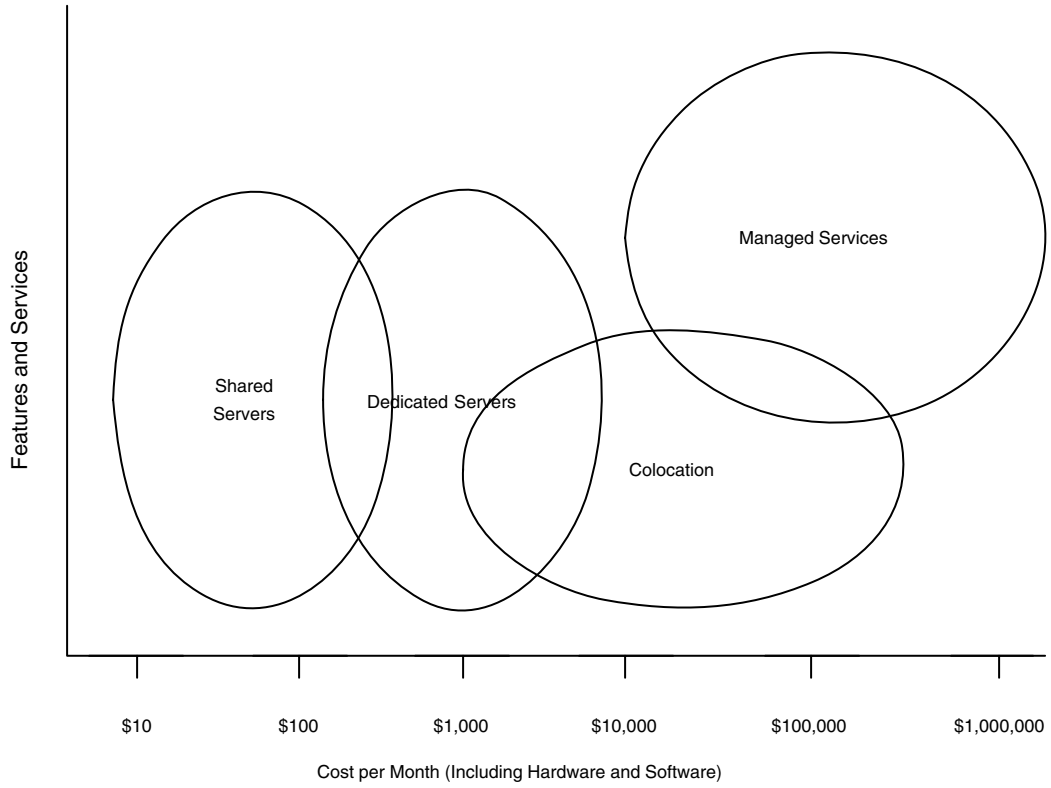


Figure 4: Cost comparison of web-hosting service categories.

excellent choice for simple, *brochureware* sites (i.e., those that contain only marketing and promotional content and don't support complex e-commerce). Major Web-hosting services that specialize in shared servers typically build large *server farms*—Internet data centers with rows upon rows of racks filled with shared servers—which is how these vendors achieve the economies of scale necessary to offer shared-server hosting at low monthly prices. But because prices are low, and profit margins are therefore slim, shared-server hosting includes very little in the way of handholding services such as help troubleshooting or configuring of a customer's Web site.

Shared-server hosting is sold in packages that typically range in cost from $20 to $500 per month. Most vendors offer more than one package and allow customers to migrate or upgrade to a larger or more expensive package as the customers' needs increase.

Although the details of shared-server packages can vary greatly, the standard means by which vendors define their packages are the following:

Monthly fee (recurring),
Setup fee (one-time or nonrecurring),
Monthly data transfer cap (maximum),
Maximum disk storage, and
Number of e-mail accounts (such as info@yourdomain.com).

Table 1 lists these parameters for three sample shared-server packages.

Typically, these packages are designed so that if users exceed the limit of any one of the measurements (data transfer, disk storage, or number of e-mail accounts), they are expected to move up to a more expensive package. Using the sample packages in Table 1, for example, if a site required 400 MB of disk storage but would only transfer 2 GB of traffic per month onto the Internet, it would be forced to buy the midrange package rather then the low-end one.

Volume and Standardization

The computer hardware on which shared Web sites run varies greatly. Some shared-server vendors use a small number of large servers, each of which can host thousands of sites on a single computer system. Increasingly, however, shared-server vendors are turning to larger numbers of less powerful, compact servers that are physically only 1U in height. (A "U" or *rack unit* is 1.75 inches.)

Table 1 Typical Shared-Server Packages (as of April 2003)

	Low-End	Midrange	High-End
Monthly Fee	$20	$75	$500
Setup Fee	$25	$75	$250
Monthly Data Transfer	10 GB	25 GB	100 GB
Disk Storage	250 MB	500 MB	1 GB
Email Accounts	30	70	150

In order to support such a large number of sites at such a low price, shared-server hosting is necessarily based on a very high degree of standardization. All components of shared-server hosting are treated as though they were part of assembly lines.

Shared-server hosting vendors typically offer menus of features and components from which a customer can select when building a Web site. Fortunately, due to intense competition in the shared-server hosting business, these menus include a large number of features, and it is relatively easy to compare vendors' offerings to one another.

The following is a typical menu of shared-server hosting features:

Daily backup to tape,
Off-site tape storage,
E-mail accounts (mailboxes),
Outbound e-mail relaying,
E-mail redirectors,
E-mail autoresponders (for automated responses to info@yoursite.com, etc.),
Microsoft FrontPage extensions,
Discussion forum software,
Anonymous FTP,
Administrative access via telnet or secure shell (SSH),
Electronic shopping cart software,
Secure Socket Layer (SSL) for secure Web pages and forms,
Credit card merchant accounts and transaction processing,
Log file processing and analysis tools,
Support for scripting languages such as Perl and PHP,
Web-based control panels or access to configuration files for managing Web sites,
Simple database software such as MySQL,
Firewall protection of the Web server, and
Streaming media servers (optional, at additional cost).

In addition to providing these site components, a shared-server hosting vendor maintains all of the hardware and software.

Three Tiers of Shared-Server Vendors

The ranks of shared-server vendors are actually divided into three different *tiers* (or subcategories), and each tier offers advantages in certain situations.

Facility-owner vendors. Vendors that own their servers and also operate their own data centers.

Tenant vendors. Vendors that own their servers, but rent space within colocation facilities.

Resellers. These vendors do not even own the hardware, but rather act as agents of facility-owner or tenant vendors.

Facility-Owner Vendors

Today, the state of the art for data centers is quite high, and very few vendors (shared-server or otherwise) have

the capitalization or borrowing power required to build and operate such facilities. On the other hand, these first-class facilities do exist, and there is no reason that a Web site—no matter how small it may be—should not reap the benefits of being located in one. Most of the owners of first-class data centers, however, do not directly offer shared-server hosting, because they are more likely to offer services at the high end of the hosting spectrum. The vendors that own large data centers are more experienced in dealing with customers who have Web-hosting budgets on the order of $10,000 per month or more.

Tenant Vendors

Many smaller Web-hosting services came to realize over time that their second-class infrastructure not only was insufficient, but was also causing them to lose customers to competitors who had better Internet connectivity, power, air conditioning, and physical security. The smart ones—those who realized they would otherwise be at a competitive disadvantage—adopted an "if-you-can't-beat-'em-join-'em" attitude and opted out of providing any of the physical aspects of Web-site hosting. They themselves are customers (or *tenants*) of colocation facilities, but because they, in turn, offer shared-server hosting, they are *tenant vendors*. A tenant vendor can be an excellent alternative for a shared-server customer. The customer gets the advantages of a first-class data center and of having someone who is focused entirely on the operation of the shared-server systems.

Resellers

Resellers are the third tier of shared-server vendors. A reseller bundles shared-server Web hosting from a third party with the reseller's own value-added services.

At first, the practice of reselling was considered deceptive, and today there are still some resellers who attempt to hide what they are doing from their customers. The better resellers, however, recognize that both they and the parent Web-hosting service are providing value to the end customer, and these resellers understand that it makes more sense for the customer to be aware of exactly what is happening.

From one perspective, resellers are simply salespeople who sign up customers and collect money. But there are some resellers who actually provide a substantial level of service above and beyond the services provided by the parent vendors for whom they resell. For example, some resellers (like some tenant vendors) provide the site design, setup, and maintenance services that a high-volume shared-server vendor does not offer.

Likewise, some resellers and tenant vendors have particular *vertical-application expertise* that may be of substantial value and importance. Vertical-application expertise means experience with shared-server hosting to specific industry groups. There might be a reseller or tenant vendor, for instance, that offered shared-server hosting for realtors. Such a vendor might include a number of services that are unique to getting a real estate brokerage site online.

Dedicated Servers

Dedicated servers are essentially the same as shared servers (i.e., they are owned by the hosting service and come with a standard suite of software and tools), but each Web site gets its own server and does not have to share it with other sites. In fact, the same companies that offer shared-server Web hosting typically offer dedicated servers as well, and some treat the two categories as virtually identical.

As compared to shared servers, dedicated servers offer the following advantages:

Increased capacity. Dedicated servers can handle substantially more of everything than a shared server can. Some dedicated servers can handle up to the following limits:
- 100 GB disk space
- 1 TB (a terabyte, or 1,000 gigabytes) monthly data transfer
- Peak data transfer rates approaching 100 Mbps

Improved security. Dedicated servers eliminate the risks associated with sharing hardware with other Web sites.

Improved reliability. A dedicated server is less subject to outages and slowdowns caused by interactions with other Web sites.

Additional configurations. Some Web-hosting services allow customers to build sites using multiple dedicated servers that can be configured in a variety of ways to increase reliability, capacity, functionality, or all three.

The cost of dedicated-server hosting ranges from $100 to $20,000 per month, and like shared-server hosting, it is sold in packages. Table 2 lists some typical examples.

Two Tiers of Dedicated-Server Vendors

Unlike the shared-server business, where such relationships are common, few resellers or tenant vendors are in

Table 2 Typical Dedicated-Server Packages (as of April 2003)

	Low-End	Midrange	High-End
Monthly Fee	$125	$1,000	$10,000
Setup Fee	$250	$1,000	$10,000
Server Hardware	(1) 1-CPU server	(2) 2-CPU servers	(3) 1-CPU servers (1) 2-CPU servers
Software	Linux, Apache	NT, IIS	Solaris, Apache
Monthly Data Transfer	200 GB	500 GB	2,000 GB
Disk Storage	8 GB	36 GB RAID	250 GB RAID

the dedicated hosting business. There are, however, two distinct types of dedicated-server vendors.

Shared/dedicated vendors. Vendors that offer both services. In most cases, their dedicated-server business grew out of their shared-server business. As their shared-server customers' requirements increased, these vendors moved those customers to dedicated hardware.

Dedicated-only vendors. Vendors that are part of a somewhat newer breed. They do not offer any shared-server hosting.

Shared/dedicated vendors tend to treat both categories of customers (shared or dedicated) alike. Because these vendors evolved from the shared model, that is how they tend to view all Web sites. Many of the issues regarding shared servers earlier in this chapter also apply to this tier of dedicated-server vendors. For example, shared/dedicated vendors are committed to standardization and the self-service model.

Because dedicated-only vendors are comparatively new, they began with a fresh start. These vendors tend to offer somewhat higher levels of customization and professional services than are available from the shared/dedicated vendors, and, not surprisingly, they tend to do so for a higher price.

Shared and shared/dedicated vendors generally offer their support services in a depersonalized manner. Typically, shared and shared/dedicated vendors build call centers into which all e-mail messages, calls, and alerts are funneled through a single queue, with the one possible exception of segregating NT and Unix/Linux issues because the skill sets necessary to support Web sites based on these two families of operating systems are so different.

Because the average revenues per customer are substantially higher for dedicated-only vendors, these vendors tend to incorporate at least some level of personalized service into their offerings.

COLOCATION

Now that we have covered shared and dedicated hosting, let us look at *colocation* (or *colo*). This is the oldest and most basic of the four Web-hosting services, and unlike shared or dedicated hosting, it is aimed at high-budget, sophisticated customers. Colocation vendors supply the fundamental services that are sometimes referred to as "power, pipe, and ping," a catchy phrase that includes, at the bare minimum, the following services:

"Real estate" (equipment racks, cabinets, or cages)
Electrical power (including battery and/or generator backup)
Air conditioning
Physical security
Fire suppression
Connectivity to the Internet

Some colocation vendors provide the following ancillary services:

Domain name service (DNS)
"Remote hands" to reboot servers or to cycle them off/on
Basic Web-site monitoring and alert notification (pagers, phone calls)
Swapping of backup tapes (i.e., the customer manages the backup, but the colocation service removes and replaces tapes from the drives)
Hardware installation services
Spares management (i.e., management of spare parts made available to hardware repair technicians)

Three types of facilities are available from colocation services: open racks, cabinets, and cages. The variations in pricing for colocation real estate are primarily based upon the type of space provided.

Open Racks

Open racks are best for sites that do not have enough servers to fill an entire rack or cabinet. Originally called "relay racks" because they held mechanical telephone relays before the advent of semiconductors, these come in standard 19- and 24-inch widths and are typically six or seven feet tall.

The charges are based on the percentage of the rack used, or the number of inches or *rack units*. For example, some vendors charge for each half or quarter rack, whereas others charge by the inch, foot, or "U." (One U or *rack unit* is equal to 1.75 inch. A server or other piece of rack-mountable equipment that is 4U in height, for instance, requires 7 inches of vertical rack space.)

Table 3 shows typical pricing for a site that has four small servers occupying one-half of a single open rack. Note that in this example, the vendor will be rotating

Table 3 Sample Partial-Rack Colocation Pricing (as of April 2003)

		One-Time (Setup)		Monthly	
Service	**Qty**	**Each**	**Total**	**Each**	**Total**
Half Open Rack	1	1000	$1,000	$ 500	$ 500
Minimum Bandwidth (1Mbps)	1		$ -	$1,000	$1,000
Remote Hands Service (Per Server)	4		$ -	$ 50	$ 200
Daily Tape Rotation (Per Tape)	1		$ -	$ 500	$ 500
Totals			$1,000		$2,200

Table 4 Sample Locked-Cabinet Colocation Pricing (as of April 2003)

		One-Time (Setup)		Monthly	
Service	**Qty**	**Each**	**Total**	**Each**	**Total**
Cabinet Rental (Per Cabinet)	2	1200	$2,400	$700	$1,400
Minimum Bandwidth (1Mbps)	2		$ -	$750	$1,500
Additional Bandwidth (Per 1 Mbps)	3		$ -	$750	$2,250
Daily Tape Rotation (Per Tape)	2		$ -	$500	$1,000
Totals			$2,400		$6,150

backup tapes and providing a basic Remote Hands service to perform simple tasks such as cycling server power off and on.

Locked Cabinets

Locked cabinets are an alternative to open racks. Not only do they reduce the number of problems that occur as a result of maintenance to neighboring systems, but they also improve security. In addition, a locked cabinet can be used as a place to keep tools and spare parts—something that cannot be done with open racks.

Charges for locked cabinets are per cabinet. The example in Table 4 is for a site using two locked cabinets and a total of 5 Mbps of bandwidth.

Cages

Finally, *cages* can be used to create small, private data centers. Data center cages look very much like those in a zoo. They are typically made from a material that resembles Cyclone fencing, but is much tougher. They are fully enclosed (i.e., they go all the way to the ceiling) and have doors with padlocks. The smallest cages are approximately 7 feet by 10 feet and can hold three or four open racks. Larger cages can hold hundreds of racks. (Open racks are not a problem when they are used inside a private cage. In fact, it is much easier to cable and maintain servers installed in open racks than when they are installed inside enclosed cabinets.)

Most Web-site owners find that cages are particularly attractive if they need more than five or six racks or cabinets and expect to manage their own servers or use an MSP independent from the colocation vendor. A cage gives the user the ultimate in isolation from other customers. And because users have floor space in addition to the rack space, they can store even more tools, spare parts, and servers than they can store in cabinets.

The charges for cages may be based on the size of the cage, the number of racks within the cage, or both. Table 5 is a sample of cage colocation pricing.

Bandwidth Costs

Nearly all colocation vendors charge for connectivity to the Internet using the 95th Percentile Rule, in which the 5% of busiest bandwidth measurements, taken every 5 minutes, are discarded, and the next-highest measurement is the one used for billing purposes. The retail cost of bandwidth at relatively low volumes begins at a maximum cost of about $1,000 per Mbps per month. For example, a site that delivers 1 million page views per month might use a 95th percentile peak bandwidth of 5 Mbps. The cost (before any quantity discounting) would be $5,000 per month.

MANAGED SERVICES

For many years, if customers opted for colocation, they had no choice but to manage their colocated servers themselves. Over time, more and more Web-site owners found themselves in this position. Because those owners tended to have the largest budgets, a new market opportunity appeared for someone who was willing and able to come in and manage the high-end Web sites housed at colocation facilities.

To exploit this opportunity, a new Web-hosting service category, managed service providers, was born. Managed services are specifically designed to work in conjunction with colocation and to provide those services that are not addressed by colocation vendors themselves.

Table 5 Sample Cage Colocation Pricing (as of April 2003)

		One-Time (Setup)		Monthly	
Service	**Qty**	**Each**	**Total**	**Each**	**Total**
Cage Rental (10′ x 10″ with 4 Racks)	1	5000	$5,000	$6,500	$6,500
Minimum Bandwidth (1 Mbps Per Rack)	4		$ -	$ 750	$3,000
Additional Bandwidth (Per 1 Mbps)	6		$ -	$ 750	$4,500
Remote Hands Service (Per Server)	14		$ -	$ 50	$ 700
Daily Tape Rotation (Per Tape)	4		$ -	$ 500	$2,000
Totals			$5,000		$16,700

Many colocation vendors now offer managed services in reaction to the success of the MSPs, some of which were acquired by the colocation vendors. But for years, although customers screamed that they needed such help, colocation vendors simply could not provide managed services themselves. The reason comes back to a recurring issue: The skills that are required to provide good power, pipe, and ping (colocation) are very different from those required to support and manage servers and applications. Even the cultures of such organizations are very different. (Just imagine calling a phone company's service center for help with a database performance problem, to understand the difference.)

MSP Segmentation

When we covered shared, dedicated, and colocation Web hosting, we were able to further break down those classifications into subcategories. But because of the relative newness of the managed service business, the breakdown of vendors and the definitions of services are not as precise or as widely accepted as with shared, dedicated, and colocation hosting. The jargon has not yet stabilized, and an extraordinarily wide range of companies now call themselves MSPs.

So rather than create categories that are not already in use in the MSP industry, our approach will be to identify the raw criteria that distinguish one MSP from another. The criteria we explore include the following:

Flexibility. Some MSPs support a very limited (rigid) set of software and hardware products, whereas others are quite flexible in this regard. But as we will see, flexibility is not necessarily a good thing.

Facility neutrality. Some MSPs own their own data centers, whereas others are *portable* or data-center independent.

Service levels and pricing models. MSPs offer different levels of service and use pricing models that range from time-and-materials to flat-rate (component) pricing.

We will examine each of these criteria in more detail and discuss how to select an MSP.

Vendor Flexibility

The first MSPs approached the task of managing Web sites in much the same way as the classical IT staff outsourcers. They simply provided a staff that was skilled in Web operations on a professional services basis and billed by the hour. Such MSPs still exist and are referred to here as *flexible* MSPs, for reasons that will become clear shortly.

Other MSPs recognized, however, that they could achieve certain economies of scale, and perhaps even offer services of superior quality, by standardizing on a specific, limited set of hardware and software products and repeatable processes and procedures to support those products. These are referred to as *rigid* MSPs ("rigid" is not a derogatory term, and it is the best word to describe this class of vendors). There are advantages and disadvantages to both the rigid and flexible models, and there are MSPs at all points in between, as illustrated in Figure 5.

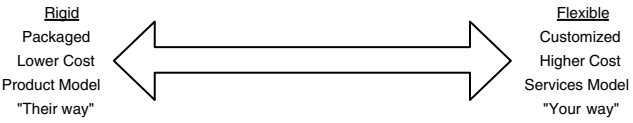


Figure 5: Rigid versus flexible MSPs.

The Flexible Model

At one end of the spectrum, fully flexible MSPs are essentially contract-staffing organizations that will support any technologies, hardware, platforms, and applications. They are in the professional services business, providing system administrators, database administrators, and others either on a full-time (dedicated) basis or on an on-call basis and shared with other clients.

Most flexible MSPs add additional value in two ways. First, they tend to specialize in certain technologies, either companywide or by employing individuals with specialized skills and experience. Second, they implement a variety of processes and systems that are shared by all of their customers. These systems include monitoring a Web site's uptime and performance, as well as customer resource management (CRM) components such as call-center and incident-tracking systems.

The general approach of a flexible MSP is to provide services *the customer's way*, and to try to function as an extension of the customer's own in-house staff. The primary advantage of working with a flexible MSP is that customers can build their Web sites using any hardware and software. If they want to use an application developed by ABC running on servers made by XYZ, no problem. If it is a combination the MSP has never encountered before, it will ramp up on it, and do whatever is necessary to make it work.

But the real benefit of using a flexible MSP shows up later, when customers decide to enhance their sites by adding new features and systems that they could not have anticipated at the time they selected their MSPs and designed their initial configurations. If an organization tends to live on the cutting edge of Web and Internet technologies, it might do well to work with an MSP that can commit to this type of flexibility.

Of course, flexibility does not come without a downside. As we will see, more rigid MSPs can sometimes achieve higher levels of efficiency and reliability due to their focus on repeatability and scalability. Although a flexible MSP can and will do whatever customers need, it may cost more in dollars, time, and reliability.

The Rigid Model

"Do one thing—but do it better than anyone else," could be the mantra for the more rigid MSPs. By limiting the number of hardware and software technologies they support, they can do a better job than if they had to spread their resources across a broader range. This is not marketing fluff, but a very real advantage of this model. By doing the same things day in and day out, and creating economies of scale, a rigid MSP can simultaneously improve quality and keep costs down.

By working with a rigid MSP that is committed to a predefined and limited set of products, the customer not only shares staff with other customers (which is the case

with flexible MSPs as well, of course), but also shares the cost of training that staff, because customers are all using the same technology. Furthermore, when a new problem occurs with a given component or combination, there is a good chance that another customer will have experienced the problem first, giving the benefit of a preventive fix, before the problem shows up on a given customer's Web site.

The rigid model is not perfect, either, of course. The obvious disadvantage is that customers often cannot have things entirely their way. If an MSP supports only Oracle databases, for instance, but a customer wants to use Sybase, it will have to go to another vendor or manage the database itself.

Facility Neutrality

The primary value of an MSP is its ability to manage Web sites and keep them running. The physical infrastructure of the data center—the air conditioning, power, and security—is not the forte of most MSPs. Yet these things are still required. Each MSP, therefore, has a way to provide such infrastructure for its customers. How they do so, however, varies greatly, and this is another important criterion for an evaluation of MSPs.

Figure 6 is a family tree of MSPs, illustrating how they can be further divided into groups according to their relationships to data center facilities.

The first group—*facility-owner* MSPs—comprises those that own their data centers. In addition to being MSPs, these vendors are also in the colocation business, and they manage Web sites and servers that are located in their own facilities.

The other MSPs all belong to the *facility-neutral* group. MSPs in this category do not own data centers. Instead, they provide management services for sites and servers that are colocated at third-party facilities.

The facility-neutral group is further split into two subgroups according to whether the MSPs work only with specific colocation vendors or are entirely independent of the hosting location. The first subgroup (*tenant MSPs*) rent dedicated space within the data centers of one or more colocation vendors and require that their customers' sites be housed at data centers operated by one of those partners. Tenant MSPs have these special relationships with one or a small number of colocation vendors.

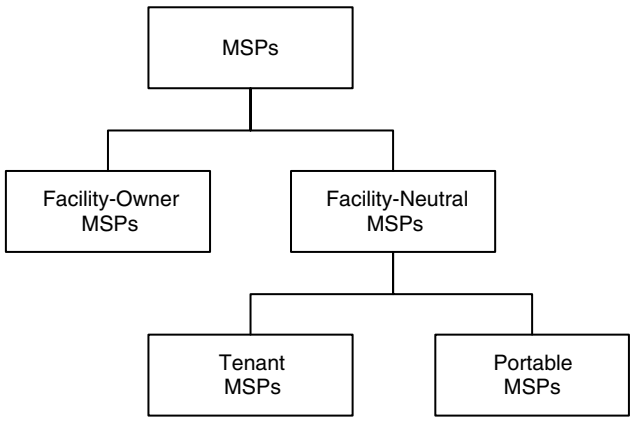


Figure 6: MSP facility ownership.

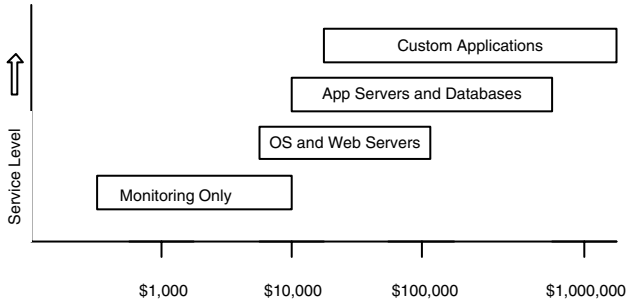


Figure 7: MSP monthly pricing ranges.

Portable MSPs, members of the other subgroup of facility-neutral MSPs, are *entirely* neutral and will manage sites located anywhere in the world, even in their customers' own corporate facilities. MSPs in this group neither own nor rent data center space.

Service Levels

The range of services offered by vendors that call themselves managed service providers and the range of fees paid by their customers have a big spread. As a group, MSPs are generally aimed toward the high end of the overall Web-hosting marketplace, with contracts starting at $1,000 per month. But some managed service contracts exceed $1 million per month—a 1000:1 range.

The range of services and the associated monthly prices are illustrated in Figure 7. The monthly fees shown are in addition to the capital expenses of hardware and software.

Monitoring

Note that at the very low end, starting at about $500 per month, MSPs provide only monitoring services. This means that for the established price (up through as much as $10,000), they will monitor a site and notify the Web site owner or specified third parties in case of outages.

The only reason MSPs offer such a basic service is to provide an entry point for customers that they hope will upgrade to higher cost services. The difference between monitoring for $500 per month and $10,000 per month depends on the number of the servers monitored and the complexity of the Web site.

Platform Management

Basic monitoring does not include services to diagnose or resolve problems. The first level of service at which MSPs take responsibility and agree to meet service levels is in the management of the *platform*. At this level, an MSP will take responsibility for infrastructure, server hardware, the operating system software, other highly generic components (such as Web-server software), and backup and recovery.

Platform management starts at about $5,000 per month. The high end ($100,000 per month) represents a site that includes 10 cabinets, at a higher ($10,000 per cabinet) price point.

Application Management

Managing applications (databases and application servers) is the goal for most MSPs, and for this, they

Table 6 Sample MSP Pricing (as of April 2003)

Service	Qty	One-Time (Setup) Each	One-Time (Setup) Total	Monthly Each	Monthly Total
Cabinet Rental (per cabinet)	2	1200	$ 2,400	$ 700	$ 1,400
Web Server SLA	4	1500	$ 6,000	$ 500	$ 2,000
Application Server SLA	2	2500	$ 5,000	$1,500	$ 3,000
Database Server SLA	1	5000	$ 5,000	$2,500	$ 2,500
Bandwidth (per Mbps)	5		$ -	$ 750	$ 3,750
Monitoring (per server)	7		$ -	$ 300	$ 2,100
Firewall Service (per public server)	4	1000	$ 4,000	$1,200	$ 4,800
Load Balancing (per public server)	4	1000	$ 4,000	$ 800	$ 3,200
Security Services (per server)	7	1500	$10,500	$ 350	$ 2,450
Tape Backup/Restore (per server)	7	750	$ 5,250	$ 50	$ 350
Tape Backup/Restore (per GB)	50		$ -	$ 50	$ 2,500
Professional Services Retainer (per hour)	20		$ -	$ 175	$ 3,500
Totals			$42,150		$31,550

hope to generate $10,000 or more per cabinet per month in revenues. The chart in Table 6 includes a low end at this price with a single-cabinet customer. A four-cabinet customer paying $15,000 per cabinet will spend $60,000 per month.

Pricing Models

Most MSPs use a model of flat-rate fees for individual components. Many vendors continue to use T& M pricing (often based on a retainer) for some aspect of their services, but by and large they have shifted most of their charges into component-based pricing.

Here are some of the components that are often priced individually on a flat-fee basis:

A monthly charge for each cabinet

A monthly charge for each physical server (varying according to the software running on it) that covers management of the operating system and supported applications

A monthly charge for tape backup based on the number of servers and the amount of data being backed up

Separate monthly charges for load-balancing and firewall services if they are provided

Monthly rental fees for hardware and software provided by the MSP

Table 6 shows a typical quotation from an MSP for a midsize (by MSP standards) Web site.

Some interesting points in this example include the following:

Most components include an initial setup, nonrecurring expense (NRE) and a monthly recurring expense. The total NRE is $42,150 and the total monthly charge will be $31,550.

This MSP uses a retainer model for pricing professional services in addition to the component-based pricing. The customer pays for a minimum of 20 hours each month at a rate of $175 per hour. That amount is due whether or not the hours are actually used during the month. Additional hours are charged at the same hourly rate.

Unfortunately, managed-service providers do not post their pricing on the Internet, so comparison shopping for price requires getting a quotation from each prospective vendor.

CONCLUSION

Although Web-hosting services vary greatly, they can be classified and compared using industry-standard criteria and categories. In trying to select a Web-hosting vendor, one can begin by determining one's requirements in terms of price, bandwidth (data transfer), disk storage, processing power, and services. Once the proper category is identified (shared, dedicated, colocation, or managed services), these same criteria can be used as the basis for identifying and evaluating specific vendors.

GLOSSARY

Colocation A bare-bones hosting service in which the vendor provides only physical space, power, air conditioning, and Internet connectivity.

Facility-owner vendors A class of Web-hosting vendors that own and operate their own data centers.

Managed service provider (MSP) A Web-hosting vendor that takes responsibility for on-going management of a Web site's hardware and software.

95th percentile rule A method for measuring bandwidth utilization in which samples are taken at regular intervals, and the highest five percent are discarded. The next-highest value is used.

Nonrecurring expenses (NRE) One-time expenses such as installation and provisioning costs.

Peak bandwidth The highest volume of data transmission used or required by a Web site.

Rack unit A standard measure of the height of equipment when rack mounted. One "U" equals 1.75 inches.

Remote hands A service in which a Web-hosting vendor will perform basic tasks such as cycling a server's power off and on.

Resellers A class of Web-hosting vendors that act as sales agents for larger vendors.

Tenant vendors A class of Web-hosting vendors that own their servers but colocate them in data centers owned by other vendors.

CROSS REFERENCES

See *Application Service Providers (ASPs); Electronic Commerce and Electronic Business; Internet Literacy; Web Quality of Service; Web Services*.

FURTHER READING

Burnham, C. (2001). *Web hosting*. Berkeley, CA: Osborne/McGraw-Hill.

Kaye, D. *The IT strategy letter*. Weekly e-mail mailing list. Subscriptions and archives available at http://www.rds.com

Kaye, D. (2002). *Strategies for Web hosting and managed services*. New York: Wiley.

Netcraft. Available at http://www.netcraft.com

Schneier, B. (2000). *Secrets and lies: Digital security in a networked world*. New York: Wiley.

Stokeley Consulting. Available at http://www.stokely.com/unix.sysadm.resources/

The Web host industry review. Available at http://www.thewhir.com

Web Quality of Service

Tarek Abdelzaher, *University of Virginia*

INTRODUCTION TO WEB QOS

The Web has become the preferred interface for a growing number of distributed applications with various demands for reliability, availability, security, privacy, timeliness, and network bandwidth. These properties are often called Web *quality of service* (QoS) dimensions. In this chapter, we first review the main components of the Web architecture and describe the protocols that govern their interaction. We then discuss how this architecture and these protocols are affected by the different considerations of achieving quality of service that emerge in the context of contemporary Web applications.

WEB ARCHITECTURE AND QOS

The first question this chapter needs to address is why Web QoS emerged as a new challenge area, and what makes this challenge important. Today, the World Wide Web is by far the largest source of Internet flows. The great majority of Internet connections use HTTP, which is the protocol that governs Web access. Improving its performance, therefore, has dramatic global effects. Efforts to improve performance come in two flavors. First, infrastructure improvements are pursued, such as realizing higher bandwidth, faster servers, and better last-mile technologies. This is largely a hardware problem that motivates development of faster processors, memory, I/O devices, and communication fabrics. Concurrently, a substantial amount of research is done to make Web performance more *predictable*. Performance is said to be predictable when its quality can be *guaranteed* in advance. Many societal and commercial forces contribute to this need. In particular, the commercialization of the Internet and the pricing of many Internet services play a significant role in elevating the idea of performance guarantees from a value-added option to a primary concern driven by contractual obligations.

In traditional commercial products, consumers have grown to take quality guarantees for granted. Vendors have contractual obligations to accept returns of defective products or products that do not perform as advertised. Similarly, paying consumers of Internet-based services will soon expect a performance guarantee or a money-back statement. Much as with other services, it will be important that clients and service providers be able to negotiate mutually acceptable quality levels in the service contract.

Quality of service can be directly tied to revenue. This relation is manifested today in several domains. For example, ISPs often sign mutual service level agreements (SLAs), which among other things describe the performance that the traffic of one ISP should receive in the network of the other and the corresponding fee. Closer to the end user, online trading services sometimes tie their commission fees to performance in executing the trades. The fee is waived for trades that are delayed by more than a certain amount of time. In view of this emphasis on performance as a contractual obligation, a significant amount of research has been spent on architectures for achieving quality of service guarantees on the Web. In the rest of this chapter, mechanisms and policies for QoS provisioning, QoS negotiation, and utility optimization will be discussed.

THE CHALLENGE OF QOS GUARANTEES

QoS attributes in general can be classified into two categories, temporal and nontemporal. Temporal attributes are those related to the passage of time. Their relation to time can be either direct (such as response time and queuing delay) or inverse (such as throughput). Examples of nontemporal attributes include security and fault tolerance. Guarantees on nontemporal attributes are, in some sense, an intrinsic function of the algorithms that provide them. For example, an encryption algorithm generally has well-defined properties such as the complexity of the encryption code. These properties directly represent the security guarantee given to the user. Temporal properties, on the other hand, are a function not only of their providing algorithm, but also of external environmental factors such as resource availability and client load. A scheduling algorithm, for example, cannot provide a guarantee on maximum queuing delay without prior knowledge of hardware speed and the worst-case client arrival pattern. Unfortunately, this information is often unavailable a priori, thereby complicating the performance guarantee problem.

A significant body of literature has addressed the issue of guaranteeing temporal QoS attributes in the absence of adequate prior knowledge of operating service conditions such as load and resource capacity. Until recently, the current state of the art in providing acceptable temporal performance to the users has been over-design. Throwing money and hardware at a performance problem eventually ensures that there are enough resources to service all incoming requests sufficiently fast. This approach, however, is inadequate for several reasons. First, it is rather expensive, because more resources are expended than is strictly necessary. Second, it provides the same service to all clients. In many cases, however, a service provider might want to use performance differentiation as a tool to entice clients to subscribe to a "better" (and more expensive) service. Third, the server provides only a best-effort service in that there are no bounds on worst-case performance. It is sometimes advantageous to be able to quantitatively state a performance guarantee for which users can be commensurately charged.

In the following sections, we describe several approaches for QoS guarantees in more detail. We begin by a brief review of the current Web architecture and the underlying principles of Web server operation. We then survey the modifications suggested to this architecture to provide performance guarantees to Web clients.

CURRENT WEB ARCHITECTURE

From an architectural standpoint, the World Wide Web is a distributed client–server system glued together by the hypertext transfer protocol (HTTP), which is simply a request–reply interface that allows clients (browsers) to download files from servers by (URL) name and allows one page to reference another, creating a logical mesh of links. The architecture is completely decentralized. By creating links from existing content, new content is seamlessly integrated with the rest of the Web.

The HTTP Protocol

The main exchange between clients and servers occurs using the HTTP protocol. When a client requests a page from a server only the text (HTML) portion is initially downloaded. If the page contains images or other embedded objects, the browser downloads them separately. At present, two important versions of HTTP are popular, namely HTTP 1.0 and HTTP 1.1. The most quoted difference of HTTP 1.1, and the primary motivation for its existence (Mogul, 1995), is its support for persistent connections. In HTTP 1.0, each browser request creates a new TCP connection. Because most Web pages are short, these connections are short-lived and are closed once the requested page is downloaded. Unfortunately, TCP is optimized for long data transfers. Each new TCP connection begins its life cycle with a connection set-up phase, followed by a slow-start phase in which connection bandwidth gradually ramps up from a low initial value to the maximum bandwidth the network can support. Unfortunately, short-lived connections, such as those of HTTP 1.0, are closed before reaching the maximum bandwidth. Hence, transfers are slower than they need to be.

Persistent connections in HTTP 1.1 avoid the above problem by sending all browser requests on the same TCP connection. The connection is reused as long as the browser is downloading additional objects from the same server. This allows TCP to reach a higher connection bandwidth. Additionally, the cost of setting up and tearing down the TCP connection is amortized across multiple transfers. The debate over which protocol is actually better is stillgoing on. For example, a disadvantage of HTTP 1.1 is its reduced concurrency, because only one TCP connection is used for all objects downloaded from the same server, instead of multiple concurrent ones. Another problem with HTTP 1.1 is that the server, having received and served a request from a client, does not know when to terminate the underlying TCP connection. Ideally, the connection must be kept alive in anticipation of subsequent future requests. However, if the client does not intend to send more requests, keeping the connection open only wastes server resources. The present default is to wait for a short period of time (around 30 s) after serving the last request on a connection. If no additional requests arrive during that period, the connection is closed. The problem with this policy is that significant server resources can be blocked waiting for future requests that may never arrive. It is therefore not obvious that the bandwidth increase of HTTP 1.1 outweighs its limitations.

Caching and Content Distribution

To improve Web access delays and reduce backbone Web traffic, caching and Web content distribution services have gradually emerged. These services attempt to redistribute content around the network backbone so that it is closer to the clients who access it. The difference between caching and content distribution lies in a data-pull versus a data-push model. Whereas caches store content locally in response to user requests, content distribution proxies proactively get copies of the content in advance.

There are generally three types of caches, namely proxy caches, client caches, and server caches. Proxy caches are typically installed by the ISPs at the interface to the network backbone. They intercept all Web requests originating from the ISP's clients and save copies of the requested pages when replies are received from the contacted servers. This process is called page caching. A request to a page that is already cached can be served directly from the proxy, thereby improving client-side latency, reducing server load, and minimizing backbone traffic for which the ISP is responsible to the backbone provider. An important question is what to do when the cache becomes full. To optimize its impact, a full cache retains only the most recently requested URLs, replacing those that have not been used the longest. This is known as the least-recently-used replacement policy. Several variations and generalizations of this policy have been proposed, e.g., to account for page size, cost of a page miss, and the importance of the client.

To improve performance further, client browsers locally cache the most recently requested pages. This cache is consulted when the page is revisited (e.g., when the client pushes the "back" button), hence obviating an extra access to the server. Finally, some server installations

use server-side caches (also known as reverse proxies) to reduce the load on the server. The caching infrastructure significantly affects the user-perceived performance of the Web. Arlitt, Friedrich, and Jin (1999) compare the effects of different replacement plocies on cache performance. Recent research efforts address developing caches that provide some form of performance differentiation or QoS guarantees. A proxy cache, for example, can offer preferential treatment to content requested by a certain subset of clients or content belonging to a certain subset of providers. This mechanism will be described in later sections.

PERFORMANCE GUARANTEES IN WEB SERVERS

In this section, we describe performance guarantee mechanisms for Web server end-systems. As a running application example, consider a Web server farm that hosts multiple Web sites on behalf of different content providers. Web hosting is a growing business in which major investments are made by companies such as Intel, IBM, and Hewlett Packard. We discuss the type of performance guarantees required, the parties to whom the guarantees are made, and the mechanisms used to enforce these guarantees. The server farm example provides a context for describing the general classes of server performance challenges and helps illustrate solutions needed when resources are shared among multiple parties with different QoS requirements.

A Web hosting service interacts with at least three different parties: (i) end users who access the hosted content, (ii) content providers who own the Web sites exported for hosting, and (iii) network providers who provide Internet connectivity to the Web hosting farm. End users are typically interested in a fast response time; content providers care more about throughput, which translates into total capacity dedicated to their Web sites; and network providers care primarily about network bandwidth consumed by the hosting installation. In general, the mechanisms to provide these guarantees lie either on servers (i.e., on the end system) or inside the network. Below, we investigate QoS mechanisms on the end system. Network-level mechanisms are described elsewhere in this encyclopedia.

Performance Isolation

The most basic guarantee needed among multiple traffic classes is that of performance isolation. Informally, the guarantee states that the performance seen by any particular class of clients should be independent of the load imposed by any other class. For example, consider a Web server that hosts two Web sites, A and B, where B is a very popular site. Unless proper action is taken, B may overload the entire server, preventing the clients of A from accessing the server. Performance isolation imposes limits that prevent a site such as B from monopolizing the server. There are several different ways performance isolation may be implemented. They typically rely on some form of resource allocation and admission control. In the following subsections, we discuss some of the most important mechanisms for performance isolation in the operating system, middleware, and application layer.

Operating Systems Solutions

The core mechanism for performance isolation is resource reservation. Each Web site must be allocated its own resource quota on the Web server. Requests for that Web site are allowed to consume only those computing resources that are within the site's quota. Excess load on one site should not be allowed to divert resources dedicated to other sites. Traditionally, resource allocation and management is the responsibility of operating systems. Hence, the most fundamental solutions to performance isolation are operating system solutions.

Generally speaking, in a shared computing system, such as a server farm shared by multiple cohosted Web sites, common resources can be categorized into two different types; those that are shared in time and those that are shared in space. Space-shared resources include, for example, memory and disk space. Different processes may own different subsets of the resource concurrently and be denied access to resources owned by others. Space-sharing implies that the resource can be partitioned. Some resources, however, are indivisible. The prime example is the CPU. Indivisible resources can only be time-shared. In other words, they can only be allocated in their entirety to one process at a time. Clients are queued up on the time-shared resource. The queuing order and duration of resource access allowed by each client decide how the resource is shared.

Traditional operating systems such as UNIX implement a time-sharing scheduling policy, which allocates the processor one quantum at a time in a round-robin fashion among the waiting processes. This policy is inadequate for performance isolation in that it does not prevent any one class of clients from monopolizing the CPU. Consider our server farm example, where it is desired to isolate requests for different sites cohosted on the same platform. The CPU capacity available to one site under the UNIX time-sharing scheduling policy is roughly proportional to the number of processes serving that site at a given time. This number is in turn proportional to the client request rate. An overloaded Web site with a large request rate generates a large number of processes to serve these requests. It can therefore monopolize the CPU.

To ensure performance isolation, many researchers have addressed the problem of reservation of time-shared resources. The first effort on this subject came from the Real-Time Mach project at Carnegie Mellon University and is called *processor capacity reserves*. The idea of processor capacity reserves is to implement separate accounting entities (the reserves), which keep track of the processing budgets allocated to abstract CPU activities. For example, a programmer can associate a budget with each separate Web site on the hosting server. The budget specifies the percentage of time that the CPU is allowed to spend on the corresponding site, e.g., 4 ms every 30 ms. The budget is replenished periodically. To enforce performance isolation, when a Web server reads a request, the request is classified and the corresponding budget is charged for the time it takes to serve the request. If the reserve is exhausted before the end of the period, the

budget is said to have expired. When a budget expires, processes charged to that budget are blocked until the next replenishment period. Hence, these processes cannot jointly exceed the CPU allocation specified for their Web site.

One limitation of the aforementioned technique is the way accounting is done in the operating system. The party that the CPU is working for at any given time is identified in the operating system as either the kernel or a particular user process. Hence, only total kernel execution time and the total execution time of any single process can be measured. In particular, kernel execution time is not properly broken into independent activities. For example, the kernel-processing time of incoming requests is not properly attributed to the site for which the requests are destined and is thus not charged properly to the correct reserve. The problem is particularly severe in monolithic operating systems such as Linux, as opposed to microkernels such as RT Mach. In monolithic operating systems the entire TCP/IP protocol stack processing is done in the kernel and is therefore not properly accounted for. A different accounting mechanism is needed.

Resource containers (Banga, Druschel, & Mogul, 1999) have recently been proposed as a new operating system abstraction for resource reservation and performance isolation in monolithic kernels running server end systems. The authors of this approach make the observation that the accounting problem arises from the fact that in traditional operating systems the resource principal (i.e., the system entity capable of owning resources) is generally associated with a protection domain (such as a process). In server end-systems, this association is not appropriate. The logical resource principal could be a user or a content provider. Depending on the server architecture, multiple processes can serve the same principal, or a single process can serve multiple principals. The resource principals should be charged for processing that occurs both on the kernel level and in user space. Banga et al. (1999) propose the abstraction of resource containers to resolve the dichotomy. In this approach, packet filters are placed at the lowest level of the protocol stack. These filters demultiplex incoming traffic into distinct categories, e.g., by IP address. All kernel and user-level processing of traffic in each category is charged to the corresponding resource container. As before, when the container is exhausted, the operating system stops processing the corresponding traffic type until the container is replenished. The approach has been shown to provide excellent isolation. It has also been shown to be very efficient in isolating denial of service attacks.

The need for early demultiplexing presents several challenges of its own. Such demultiplexing is typically performed based on fields in the TCP or IP headers. Unfortunately, in many Web applications, client classification is a little more complex (Menasce, Almeida, Fonseca, & Mendes, 2000). For example, if classification is based on accessed content, the operating system will have to peek into the HTTP header, which is an application-layer header, to determine the accessed URL. This clearly violates the modularity of the layered network architecture and requires the kernel to be aware of application-specific URL names. A more difficult case is that of a Web server hosting an e-shopping site, which may wish to classify clients based on the contents of their shopping carts. A widely quoted example of this policy is giving clients with more expensive purchases higher priority at checkout to ensure successful completion of their purchase transactions. In this case, the classification information necessary for identifying client priority is not available to the operating system, because the contents of the shopping cart are an application-specific data structure. Designing general-purpose mechanisms to express such classification constraints at the operating system level is a challenging task, unless one is willing to sacrifice the application-independent nature of general-purpose operating systems.

Another complication is that packets arriving at the bottom of the protocol stack (e.g., IP fragments) may not always have an application-layer header, which contains the information necessary for classification. These headers would be reconstructed higher in the communication protocol stack, which makes early classification more challenging. Construction of efficient packet filters for early classification based on application-specific information is therefore an important research challenge.

Middleware Approaches

When client classes are defined in an application-specific manner, a different solution to performance isolation is to develop middleware that augments generic operating system support with application-specific middleware policies. The operating system approach, described in the previous section, is perhaps the most efficient approach for fine-grained performance isolation. However, it suffers from the lack of generality and the lack of portability. The former refers to the difficulty of incorporating application-specific classification policies into the kernel. The latter (i.e., portability) stems from the fact that support for time-multiplexed resource reservation described above is still far from being a standard operating system feature. Thus, performance isolation solutions that do not *require* this support in the operating system are often preferred. These considerations lead to middleware solutions.

Middleware refers to any software layer that runs below the application and above the operating system. Generally, there are two types of middleware: that which is transparent to the application and that which requires application modification to adhere to a new interface. The former is more general in that it does not require access to application source code. Hence, a hosting service, for example, can develop in-house middleware components even when the source code for the Web server and the operating system is not available. Similarly, a middleware vendor can develop its software independently to interoperate with multiple Web server and operating system products without the need to modify their code.

An important challenge in designing transparent middleware services is to architect their (transparent) interaction with the server software. Such interaction is usually implemented by instrumenting standard dynamic shared libraries used by the server. Dynamic shared libraries are those loaded by the server at run time, as opposed to being precompiled together with server code. This feature is

made possible due to late binding, supported by most operating systems today. In late binding, references to called library functions are not resolved until the call is made at run time and the corresponding library is dynamically loaded into the server's memory. It is therefore possible to make changes to the shared library without having to recompile or relink the server software. Once the server makes the standard library call, the new (modified) library gets invoked. In the case of middleware for performance isolation, the modified shared library may implement accounting functions that approximate resource containers.

One of the most obvious libraries to instrument is the socket library. Hewlett Packard Labs researchers (Bhatti & Friedrich, 1999) were the first to suggest architectures where the socket library is replaced with a QoS-sensitive version, which implements performance isolation. In the context of regular socket calls, the QoS-sensitive library dequeues service requests from the server's well-known port and classifies them into per-class queues. The accept() or read() socket calls are modified so that no connections are accepted from a given per-class queue unless the budget of the corresponding class (maintained by the modified library) is nonzero. The scheme implements approximate performance isolation. It has been successfully integrated into a server platform sold by Hewlett Packard, called WebQoS.

Application-Layer Mechanisms

Current state-of-the-art Web servers, such as Apache (the most widespread Web server today), maintain a single process pool for all incoming requests. The single-pool architecture significantly complicates QoS provisioning because all requests are treated alike in the server. In a multi–class server, attainment of performance isolation can be significantly simplified if the server is designed with QoS guarantees in mind. The single main feature that provides most impact in that regard is to maintain a separate pool of processes for each traffic class. Once the server identifies an incoming request as belonging to a particular class, it is queued up for the corresponding process pool. Several examples of this architecture have been proposed in the literature. QoS provisioning reduces to controlling the resource allocation of each separate pool.

Service Differentiation

An important second category of QoS guarantees is service differentiation. The goal of performance isolation, discussed above, is to logically *partition* the resource so that each class of clients would get its own independent portion. Competition among classes is *eliminated* by giving each exclusive ownership over a subset of resources. In contrast, service differentiation policies do not attempt to partition the resource. The resource is *shared*. When the resource is in demand by multiple classes, the differentiation policy *resolves* the competition, typically in a way that favors some class over others. Note that performance isolation can also lead to different performance levels for different classes and hence can be thought of as a special case of "service differentiation." One main difference is that in performance isolation no resource sharing occurs. Service differentiation policies are classified depending on what it means to favor a particular class. There are several ways "favor" can be defined. In the following, we describe the most common examples and their supporting mechanisms.

Prioritization

The simplest method to provide differentiation is prioritization. Consider a situation where a Web service is accessed by two classes of clients: paying customers and nonmembers. In contemporary Web services, paying customers are usually allowed access to protected parts of the Web site that are inaccessible to nonpaying users. This type of differentiation fails to achieve its goal when the Web site is overloaded. In such a situation, a large group of nonpaying users can increase load on the server to the extent that all users (including the paying ones) have difficulty accessing the site content. Performance isolation can be applied between paying and nonpaying users, but it suffers the problem of having to decide on the relative sizes of the respective resource partitions, which typically depend on the current load. One approach to circumventing this problem is to serve clients in absolute priority order. In this scheme all client requests are queued up in a single priority queue for server access. Under overload, the queue overflows. Clients at the tail of the queue are dropped. These clients, by construction of the queuing policy, are the lower priority ones.

The problem with prioritization alone is that it fails to provide meaningful performance guarantees to clients. The top priority class receive the best service, but very little can be predicted about the performance received by other classes. Prioritization, however, becomes an extremely useful analyzable tool once combined with other techniques discussed below.

Absolute Delay Guarantees

Prioritization, in conjunction with admission control, allows customizable absolute delay guarantees to an arbitrary number of client classes. Consider a case where there are N classes of paying clients. To recover a fee, the server is contractually obligated to serve each class within a maximum time delay specified in the corresponding QoS contract signed with that class. For example, in an online trading server, first-class clients may be guaranteed a maximum response time of 2 s, whereas economy clients are guaranteed a maximum response time of 10 s. Failure to execute the trade within the guaranteed response time results in a commission waiver. Alternatively, in a Web hosting service, the content provider of each hosted site might have a QoS contract with the host that specifies a target service time and a fee paid to the host per request served within the agreed-upon delay. Hence, an overloaded hosting server, which consistently fails to meet the delay constraints, will recover no revenue from the content providers.

Because, in these examples, a host does not derive revenue from requests that miss their deadlines, admission control may be used against clients who are unlikely to met their timing constraints. The rationale for such admission control is that scarce server capacity should not be wasted on clients who are unable to make revenue. Although theoretically admission control refers to a

choice between acceptance and rejection, it is more realistic to choose between acceptance and *background service*. In other words, for the purposes of the following discussion, a rejected client is put in a separate lowest-priority queue to be served if resources permit.

Admission control has received much attention in Web QoS literature. Admission control algorithms may be classified into optimistic and pessimistic. The former type may admit clients optimistically even when they may miss their deadlines. The latter may reject them unnecessarily. Note that absence of admission control can be thought of as the extreme of optimistic admission control tests. Recent results (Abdelzaher & Lu, 2001) have shown that in a server with randomly arriving requests it is possible to use a constant-time admission control test to distinguish clients who will meet their deadlines from those who may not. The test is based on a running counter, which maintains a utilization-like metric. The counter is updated by a constant-time operation upon request arrivals and departures. If a request arrives while the counter is below a certain high-water mark, it is guaranteed to meet its deadline. Otherwise, the deadline may be missed and the request is served at the lowest priority level. Recent evaluation results of this admission control algorithm show that it rarely errs in the sense of unnecessarily rejecting requests. The test is shown to improve revenue at overload not only by eliminating resource consumption wasted on requests that miss their deadlines but also by favoring smaller requests (all other factors being equal) and hence improving server throughput. The derivation of the high-water mark representing the client admission threshold assumes that clients are served in a priority order such that more urgent requests are served first. A generalization of this test has later been proposed for FIFO scheduling.

When priority-driven scheduling is used to meet deadlines, priority should be set proportional to urgency. There are two ways urgency can be defined. In the first, clients with shorter per-class response times are considered more urgent. The resulting priority assignment is called deadline-monotonic. In the second, urgency is defined by absolute deadline. The resulting priority assignment is called earliest-deadline-first (EDF). For example, consider a request arriving at time 0 with a maximum response time constraint of 10 s. At time $t = 9$, a second request arrives of a different class with a maximum response time constraints of 2 s. According to the deadline-monotonic priority assignment, the second request should receive higher priority because its maximum response time constraint, 2, is tighter than 10. According to EDF the first request should receive higher priority because its absolute deadline is $0 + 10 = 10$, which is before the absolute deadline of the second request, $9 + 2 = 11$.

EDF has been proved to be the optimal priority-driven scheduling policy. Deadline-monotonic scheduling is optimal among time-independent scheduling policies, i.e., those where priorities are assigned independent of absolute request arrival times. Optimality, here, is defined in terms of the ability to meet a larger number of deadlines. EDF can meet deadlines when deadline-monotonic scheduling fails because it takes arrival times into account. For instance, in the above example, if the first request has an execution time of 9.5 s, and the second request has an execution time of 1.5 s, both requests meet their deadlines under EDF, but not under deadline-monotonic scheduling. A problem with EDF is that it is less commonly implemented on standard operating systems, with the exception of embedded real-time operating systems where application timing is of prime importance. Deadline-monotonic scheduling is therefore a good compromise.

Statistical Delay Guarantees

An entirely different line of reasoning is to provide statistical guarantees on delays and deadline misses. Statistical guarantees require queuing analysis of the Web server. This analysis makes two types of assumptions. First, it must make assumptions regarding the queuing structure of the Web server. Second, it must make assumptions regarding the request arrival process. In the following, we outline the most important challenges underlying the statistical approach to QoS guarantees.

Consider the first challenge, namely, deriving the queuing structure of the Web server. This queuing structure depends on the protocol used. We describe HTTP 1.0 for illustration. Consider a typical multithreaded (or multiprocess) Web server. Packets arrive from the network, causing hardware and software interrupts that, respectively, read the packets from network interface cards and deposit them into a kernel-level input queue called the *IP queue*. In between interrupts, packets in the IP queue are processed by the kernel and queued for the particular application ports for which they are destined. These ports are often called sockets. Each socket is associated with an input queue, called the *listen queue,* where incoming connection requests are queued. Independently schedulable entities in the Web server, called *worker threads,* are blocked for data to be deposited at the listen queue. These threads are unblocked by request arrivals to execute the arriving requests. Each thread implements a loop that processes each incoming request and generates a response. Worker threads that have been unblocked by request arrival become runnable.

Multiple runnable threads may exist at a given time. The order in which such threads get the CPU to execute a request is determined by the CPU scheduling policy. This policy maintains a priority queue called the *ready queue*. The thread at the top of this queue executes for a particular time quantum or until it is blocked. Request processing by a worker thread typically entails access to one or more auxiliary server resources, the most notable being disk I/O. For example, in a Web server, disk I/O is usually needed to read the requested Web page from disk. Access to auxiliary resources blocks the calling thread, at which time it is queued for I/O until the awaited resource becomes available. Each resource usually has a queue, which determines the order in which accesses are served. We call it the *I/O queue*. The resource is made available to the thread at the top of the queue, at which time the corresponding thread becomes runnable again and re-enters the CPU ready queue. When request processing is done, the worker thread sends a response back to the client. Sending the response entails queuing data into the *outgoing packet queue* for transmission on the network.

The above discussion identifies five different queues involved in the Web server's queuing structure: namely, the IP queue, the listen queue, the ready queue, the I/O queue (assuming a single I/O resource, e.g., disk), and the outgoing packet queue. The interconnection of these queues creates a queuing network with loops and other dependencies. For example, threads that repeatedly read and process disk data essentially loop between the ready queue and the I/O queue. Moreover, when the number of threads is fixed, dequeuing from the listen queue is synchronized with progress in the ready queue in that new requests are dequeued from the former only when some runnable thread has finished serving a request or has blocked. These factors make accurate analysis of the server's queuing structure difficult. Instead, many approximations are possible. For example, it is possible to consider only the queue for the bottleneck resource. The general idea is that a Web request is likely to spend most of its time waiting in the bottleneck queue.

The second challenge in providing statistical guarantees in Web servers is to identify the stochastic nature of the arrival process. Many queuing theory results assume a continuous Poisson arrival process. This process is characterized by an exponential distribution of interarrival times. This assumption does not hold for Web traffic. Research on Web characterization has identified that arrival of Web requests is generally modeled by a heavy tailed distribution (Crovella & Bestavros, 1997). One distribution that is commonly used to model request interarrival times and request execution times is the Pareto distribution. Breslau, Cao, Fan, Phillips, and Shenker (1999) also determined that URL popularity follows a Zipf-like distribution. This information is important for studies of Web performance because it helps quantify the effects of caching. To experiment with Web performance in laboratory testbeds, Barford and Crovella (1998) developed a synthetic Web workload generator that faithfully reproduces the characteristics of realistic Web traffic, including its heavy-tailed distribution, URL popularity, and reference locality characteristics. The problem of providing queuing-theoretic performance predictions for such Web traffic arriving at a server modeled by the queuing structure outlined above is still an open research topic.

Relative Guarantees

From queuing theory we know that delay experienced by a request is a function of server load. Ultimately, the only way one can reduce delay to meet deadline guarantees is to keep the load low enough. This implies denying service to some requests under high load to improve performance. In many cases, however, it is preferred that *all* clients receive service when capacity permits. One QoS-provisioning paradigm that subscribes to this model is *proportional relative differentiated services*.

Proportional relative differentiation was first proposed in the networking community in the context of delay differentiation in routers. It was since extended to server end systems. In the relative differentiated services model, it is desired that the ratio between the performance levels of different classes of traffic be fixed; e.g., it may be that the delays of two traffic classes in a network router should be fixed at a ratio of 3:1. In general, if there are multiple traffic classes in the system, and if H_i is the measured performance of class i, the relative guarantee specifies that $H_1 : H_2 : \ldots : H_n = C_1 : C_2 : \ldots : C_n$, where C_i is the weight of class i. Hence, only relative delay between any pair of classes is specified. The absolute delay can take any value. When the system approaches overload, the delay of all classes increases, although some classes see a larger delay increase than others in accordance with the relative delay guarantee. At present, mechanisms for providing relative delay differentiation are well understood, but not yet deployed.

Relative delay guarantees make most sense under moderate server load. When the load is light, all classes see identically good service (no delay). When the load is very high, all classes suffer unacceptable delays and timeouts. Hence, it is often useful to combine relative and absolute delay guarantees in a single framework. The combined architecture bounds the maximum delay of a class in addition to providing the correct delay ratio when the bound has not been reached. The architecture allows specifying a partial order on absolute and relative time constraints that determines which constraints should be relaxed first when network conditions makes the combination unrealizable.

Convergence Guarantees

In an environment where unpredictable traffic conditions make it impossible to satisfy absolute constraints, an alternative type of performance guarantees has recently been defined. This guarantee views QoS provisioning as a convergence problem and employs control theory to ensure stability and timeliness of convergence of system performance to the right specification (Abdelzaher, Shin, & Bhatti, 2002). The statement of the guarantee is that a performance metric, R, will converge within a specified exponentially decaying envelope to a fixed value, called the *set point,* and that the maximum deviation from the set point will be bounded at all times.

The absolute convergence guarantee is translated into a control loop such as those used in industrial control plants. The loop samples the measured performance metric (e.g., delay), compares it to the set point, and uses the difference to induce changes in resource allocation and load. Typically, it performs admission control to reduce load, and reallocates resources to alleviate the bottlenecks.

The use of control theory to control the performance of software processes has gained much popularity in recent years. Traditionally, control theory was used to model and control industrial processes described by difference equations. The intuitive reason that such a theory would be applicable to server performance control is that input load on a server resembles input flow into a water tank. The fill level of the server queue resembles the tank fill level. Admission control resembles a valve on the input flow pipe. Hence, the delay dynamics of a Web server are similar to flow and level control dynamics in a physical plant. The latter are well understood and can be described by difference equations. The second reason that control theory is becoming popular for server control is that feedback control loops are very robust to modeling errors. Hence, accurate models of software dynamics are not needed.

In the context of time-related performance metrics, it is interesting to classify the convergence guarantee loops depending on the performance variable being controlled. As is the case with physical plants, the controlled output of interest affects the model of the system and whether the control loop is linear or not. Because most of control theory was developed for linear processes, the ability to satisfy the convergence guarantee is often contingent on the ability to approximate the server well by a linear model.

To a first approximation, rate metrics and queue length are easiest to control because they result in linear feedback loops. Rate can be controlled in much the same way physical flow is controlled in pipes. Queue length is simply the integral of rate, and therefore is also linear. Delay guarantees are more difficult to provide. This is because delay is inversely proportional to flow. If a request arrives at a queue of length Q, with a dequeueing rate of r, the queuing delay, d, of the request is $d = Q/r$. The inverse relation between the manipulated variable (rate) and the delay makes the control loop nonlinear. At present, providing convergence guarantees on delay remains an active research topic.

Challenges

In the previous section we outlined the semantics of the most important types of performance guarantees. Next we describe challenges that are common to achieving these guarantees in Web servers.

Admission Control

A common enforcement mechanism of many QoS guarantee types in Web servers is client admission control. An important decision in the design of an admission controller is to choose the entity being admitted or rejected. In a Web server, admission control can operate on individual requests, individual TCP connections, or individual client sessions.

Per-request admission control is the simplest to implement, but has serious limitations. Consider an overloaded server operating at 300% capacity. Statistically, this means that two out of three requests must be rejected on average. For simplicity assume that all clients belong to the same class. Because client browsers typically issue multiple requests over the duration of the client's interaction with the server (even when downloading a single page), per-request admission control will uniformly cause each client to encounter failures in two thirds of the accesses. Such service is virtually unusable. Per-request admission control discriminates against longer sessions. This has very negative implications, especially from an e-commerce perspective. For example, it has been shown in many studies that those e-shoppers who eventually complete a purchase from an online server typically have longer sessions with the server than occasional visitors who do not buy. Hence, discriminating against longer sessions gives a lower QoS precisely to those users who are more likely to generate revenue.

A better admission control scheme is to select a consistent subset of clients to admit. Those clients are admitted as long as they continue to send requests to the server within a specified time interval. The rest are consistently rejected. This scheme succeeds in making the service usable by at least some consistent subset of clients. This scheme is commonly called *session-based admission control*. It was first analyzed at length for Web servers by Cherkasova and Phaal (1999) and continues to be an important research topic (Chen & Mohapatra, 2002). Session-based admission control is more difficult than per-request admission control because at the time a session is admitted the server has no knowledge of the future load that it may impose. Hence, it is difficult to decide how many sessions can be admitted before server capacity is exceeded. Different clients can impose substantially different load demands. For example, if the admitted client is a Web crawler, it may impose a much larger load than a human user. This difficulty is usually resolved by adding feedback into the admission control loop. The admission controller must continuously refine the subset of clients to be admitted based on measurements of resulting load. Feedback-based admission control schemes work very well in practice.

Client Authentication

An issue related to admission control is that of client identification. The simplest scheme is to identify the clients by their IP addresses. This identification, however, is not accurate. In several cases, the IP address of the client is unavailable to the server. For example, it is possible that the client is behind a firewall or a proxy. In this case, it is the proxy's IP address that is seen by the server, not the client's.

To identify clients, Web servers may use "cookies," which work similarly to passwords. The server sends a cookie to a client upon the first access. The client's browser automatically presents the server with this cookie when subsequent accesses are made. The server is therefore able to identify accesses that belong to the same client and separate them from those of another. This capability may be used to implement session-based admission control. In general, cookies have several security flaws. For example, it is possible for cookies to be copied, allowing a third party to impersonate the client. Moreover, cookies are sent in plain text. A better mechanism would be to encrypt the transfer. Such encryption may be provided at a lower level such as the secure socket layer.

Rejection Cost

Another issue tightly related to admission control is the cost of rejecting an incoming request. If admission control is based on the client identity obtained from browser-supplied cookies, a connection cannot be classified early inside the kernel. Instead, all requests have to reach the application layer, where the server process can interpret the cookies and identify them as belonging to a particular class. Now imagine an overloaded server, which decides to admit all requests of class A and reject all requests of class B. Because the kernel cannot tell the two classes apart, all requests are forwarded to the application layer after being processed in the kernel. Kernel processing, as mentioned before, takes a nonnegligible overhead. This overhead is incurred whether the request is accepted or not. In particular, it is incurred for each rejected request. It is

therefore called *rejection cost*. The rejection cost of a request can be more than half the cost of processing the request successfully. Hence, at overload, a significant portion of server capacity is wasted on request rejection.

Note, in comparison, that a best-effort server, which does not need to classify requests, incurs a lower cost per failed request at overload. This is because when such a server gets overloaded, the socket queue overflows in the kernel. Subsequent requests fail to get enqueued in the listen queue and are dropped much earlier in the protocol stack, hence incurring a lower rejection cost. QoS-aware servers ensure that indiscriminate tail dropping does not occur. For example, a high-priority thread is often dedicated to dequeuing the listen queue and classifying the requests, thereby increasing the cost of rejection. Minimizing rejection cost in QoS-aware servers with complex request classification policies is an important research topic.

Consistent Prioritization

Many guarantee types, such as absolute delay guarantees, usually rely on priority-driven scheduling. Prioritization imposes a significant challenge in most mainstream operating systems. To be effective, all resource queues should be *identically* prioritized. Unfortunately, CPU priorities, which can be set explicitly in many operating systems, control only the order of the ready queue. It has been shown in recent studies that this queue is often not the bottleneck. In a previous section, we have identified at least five resource queues involved in a Web server. In many cases, the largest queue in the server is the listen queue on the server's well-known port. This queue is maintained in the TCP layer and is handled in FIFO order. Correct prioritization would imply prioritizing the socket listen queues as well. In I/O intensive servers, such as those serving dynamically generated content, the I/O queue may be the bottleneck. Hence, disk access should be prioritized. Moreover, in a server implementing data structures protected by semaphores, it must be ensured that processes queued on a semaphore are awakened in consistent priority order.

Communicating priority information among multiple resources is a nontrivial undertaking. Proper operating system support must exist for priority inheritance across different resources. This support is complicated by the fact that blocking over nonpreemptive resources may cause involuntary priority inversion. The classical example of that is the case of two requests, A and B, where A is of higher priority. Let request B arrive first at some server and be blocked on a nonpreemptive resource such as a shared data structure protected by a semaphore. Request A arrives later and is blocked waiting for B to release the lock. Meanwhile, the progress of B may be interrupted by an arbitrary number of requests of intermediate priority. In this scenario, A is forced to wait for an arbitrary number of lower priority requests, even when all resource queues (including the semaphore queue) are correctly prioritized. The problem may be solved by the *priority ceiling protocol* developed at CMU, which bounds priority inversion. Unfortunately, current mainstream operating systems neither enforce resource priorities nor implement mechanisms for bounding priority inversion, such as the priority ceiling protocol. Thus, the current state of deployment is far from adequate for the purposes of implementing priority-based QoS support on Web server platforms.

Automated Profiling and Capacity Planning

In many cases, providing QoS guarantees requires developing a service execution model that describes server capacity in units of contracted work. This problem is generally called capacity planning. For example, a content provider may wish to make an agreement with a hosting server to host the business Web site of the former. The content provider may agree to pay for an expected client access rate of 100 requests/second on static content of an average size of 10 KB/request. The host contractually agrees to serve that rate. The problem of the host is to determine how much server capacity should be allocated to this site so that the contractual service obligations are met. This, in turn, requires knowing the execution overhead per request received and per byte sent of the response. A common approximation of service time of a request for static content is $time = A + B\,x$, where *time* is the service time, A is a fixed per-request overhead associated with protocol processing, x is the size of the response, and B is the overhead per unit of response data sent.

The problem with computing such execution overheads is that they depend on the hardware and software of the underlying platform. Thus, they need to be recomputed upon every platform or software upgrade. The cost of recomputing these parameters may be excessive. Fortunately, it can be reduced using automated profiling middleware. Automated profiling middleware transparently instruments the server to measure various overheads during normal operation. These overheads are then correlated with measured load (such as the measured request rate and response bandwidth) to yield the best value of execution parameters A and B. Least squares estimation is a particularly useful tool to perform such correlation. Automated profiling eliminates manual profiling costs, hence making it feasible to do accurate capacity planning in QoS-aware Web services. Techniques for accurate and robust automated profiling are currently under investigation.

QoS Adaptation

The forgoing discussion focused on controlling load to provide time-related guarantees. The underlying assumption is that service must be provided by the deadline. There are no intermediate compromises. In the following, we present a case for QoS adaptation algorithms, which can negotiate intermediate performance levels within a predefined range deemed acceptable by the user. We describe mechanisms that implement adaptation in Web servers.

The Case for QoS Adaptation

Most QoS-sensitive applications have a certain degree of flexibility in terms of resource requirements. For example, JPEG images can be adapted to bandwidth limitations by lossy compression or resolution reduction. Dynamically generated pages can be replaced by approximate static

versions to save execution time. Thus, when the server is overloaded, an alternative to rejection of further requests would be to adapt a quality of responses such that the load on the server is reduced. Many leading news and sports Web sites adopt this policy. For example, the appearance of the Cable News Network Web site at http://www.cnn.com is often significantly simplified upon important breaking news to absorb the higher request rate. An instance of that was the great reduction in CNN site content during the first hours after the attack on the World Trade Center on September 11, 2001.

Content degradation is preferred to service outage for obvious reasons. One is that it maintains service to all clients, albeit at a degraded quality, which is preferred to interruption of service. Another is that it does not incur rejection cost because service is not denied. As mentioned before, rejection cost can be considerable when user-level admission control is used.

To express the flexibility of adaptive Web applications, an expanded QoS-contract model is proposed. It assumes that the service exports multiple QoS levels with different resource requirements and utility to the user. The lowest level, by default, corresponds to request rejection. Its resource requirements are equal to the rejection cost, and it has no utility to the user. The objective is to choose a QoS level delivered to each user class such that utility is maximized under resource constraints. Several content adaptation architectures have been proposed in Web QoS literature. They can be roughly classified into two general types, depending on the reason for adaptation, namely, adaptation to network/client limitations and adaptation to server load. These two types are described below.

Adaptation to Network and Client Limitations

In the first type, adaptation is performed online and is sometimes called dynamic distillation or online transcoding. For example, see the work of Chandra, Ellis, and Vahdat (2000). The reason for such adaptation is to cope with reduced network bandwidth, or client-side limitations. Note that the dynamic distillation algorithm itself will in fact increase the load on the server. In effect, the algorithm implements a trade-off where extra computing capacity on the server is used to compress content on the fly to conform to reduced network bandwidth. Alternatively, transcoding or distillation proxies may be introduced into the network. For example, a transcoding proxy can identify a client as a wireless PDA device and convert a requested HTML page into WML for display on the client's limited screen.

Adaptation to client-side limitations can also be done using layered services. In this paradigm, content delivery is broken into multiple layers. The first has very limited bandwidth requirements and produces a rough version of the content. Subsequent layers refine the content iteratively, each requiring progressively more resources. JPEG images, for example, can be delivered in this manner. An adaptive service could control the number of layers delivered to a client depending on the client's available bandwidth. A client with a limited bandwidth may receive a fraction of the layers only. The determination of the number of layers to send to the client can be done either by the server of by the client itself. For example, consider an online video presentation being multicast to the participants of a conference call. The server encodes the transmitted video into multiple layers and creates a multicast group for each layer. Each client then subscribes to receive a fraction of the layers as permitted by its resource capacity and network connectivity. Such adaptation architectures have initially been proposed in the context of streaming media.

Adaptation to Server Load

In the second type of adaptation, content is adapted to reduce server load. In this case, dynamic distillation or compression cannot be used because the server itself is the bottleneck. Instead, content must be preprocessed a priori. At run time, the server merely chooses which version to send out to which client. The server in such an architecture has multiple content trees, each of a different quality. For example, it can have a full content tree, a reduced content tree where some decorative icons, backgrounds, and long images have been stripped, and a low-quality text-only tree. A transparent middleware solution has been described that features a software layer interposed between the server processes and the communication subsystem. The layer has access to the HTTP requests received by the server and the responses sent. It intercepts each request and prepends the requested URL name with the name of the "right" content tree from which it should be served in accordance with load conditions. To decide on the "right" content tree for each client the interposed content adaptation layer measures the current degree of server utilization and decides on the extent of adaptation that will prevent underutilization or overload.

An interesting question is whether or not load can be adapted in a continuous range when only a finite small number of different content versions (trees) are available. Such continuous adaptation is possible when the number of clients is large. To illustrate this point, consider a server with M discrete service levels (e.g., content trees), where M is a small integer. These levels are numbered $1, \ldots, M$ from lowest quality to highest quality. The level 0 is added to denote the special case of request rejection. The admission control algorithm is generalized, so that instead of making a binary decision, it determines a continuous value m, in the range $[0,M]$, which we call the degree of degradation. This value modulates server load in a continuous range. In this case, $m = 0$ means rejecting all requests (minimum load), and $m = M$ means serving all requests at the highest quality (maximum load). In general, when m happens to be an integer, it uniquely determines the service level (i.e., tree) to be offered to all clients. If m is a fractional number, composed of an integral part I and a fraction F (such that $m = I + F$), the two integers nearest to m (namely, I and $I + 1$) determine the two most appropriate service levels at which clients must be served. The fractional part F determines the fraction of clients served at each of the two levels. In effect, m is interpreted to mean that a fraction $1 - F$ of clients must be served at level I, and a fraction F at level $I + 1$. The policy can be accurately implemented when the number of clients is large. It ensures that load can be controlled in a continuous range

by fractional degradation and offers fine-grained control of delay and server utilization.

PERFORMANCE CONSIDERATIONS IN WEB PROXY SERVERS

Before this chapter concludes, a word on performance considerations in proxy servers is in order. Proxy servers are intermediaries between the clients and the accessed Web servers. They are the main performance acceleration mechanism on the Web, which makes their study very important. QoS architectures should consider the effects proxy servers have on user-perceived Web performance and make use of them to satisfy client QoS requirements. Proxies may be used for caching, transcoding, or content distribution. A proxy intercepts incoming requests and attempts to serve them locally. If the requested content is not locally available the proxy may forward the request to another server (e.g., content distribution proxies), contact the origin server and save the response (Web proxy caches), or contact the origin server, transcode the response, and forward it to the client (transcoding proxy). Although current proxy servers typically treat all clients alike, there has been much talk on making them QoS-aware. For example, the server may offer preferential treatment to some classes of clients or classes of content.

To illustrate this point, consider a content distribution network composed of multiple proxy servers situated around the Internet backbone. The distribution provider may make agreements with content providers to distribute their content preferentially for a corresponding fee. Alternatively, the distribution provider may make agreements with certain ISPs to improve the quality of service to their clients by virtue of the content distribution network. An example of such a network is that introduced by Akamai.

Several research efforts have looked at biased replacement policies in proxy caches (Kelly, Chan, Jamin, & Mackie-Mason, 1999). Such policies attempt to maximize a weighted hit ratio, where weights are set in accordance with content importance. For example, content fetched from the preferred providers can have a higher weight and therefore a lower likelihood of being replaced. Another research direction is to determine dynamically the disk space allocation of a cache or a content distribution proxy such that content of preferred providers receives a higher hit ratio. In this approach, the "performance distance" between different content types can be controlled (Lu, Saxena, & Abdelzaher, 2001). For example, one can specify that preferred content is to receive twice the hit ratio of regular content. The underlying adaptive disk allocation policy uses feedback control to translate this specification into a dynamic disk space allocation that satisfies the specified requirement in the presence of dynamically changing load patterns.

In general, with the proliferation of caching proxies and content distribution servers, it is becoming imperative to integrate server solutions with proxy solutions in a comprehensive end-to-end framework. Server and cache cooperation can implement innovative mechanisms for performance improvement. Generalization of such architectures to the area of QoS is currently an active area of research.

CONCLUSION AND FUTURE TRENDS

In this chapter, we briefly introduced the most important issues and mechanisms for providing quality of service in the modern Web architecture. The topic of providing quality of service guarantees is becoming increasingly important with the pricing of Internet services, and with the tendency to include performance requirements within the contractual obligations of service providers. In the aggregate, QoS-assurance mechanisms impose new, significantly different challenges on the research community, calling for novel theoretical foundations for delivering guarantees in an unpredictable environment.

At present, three underlying theoretical foundations are identified, whose composition may help understand the dynamics of the Web enough to provide the needed assurances. These foundations are real-time scheduling theory, queuing theory, and feedback control theory. Real-time scheduling theory has traditionally been concerned with developing scheduling policies and resource allocation disciplines that ensure satisfaction of timing guarantees in closed embedded systems. In isolation, this theory requires making strict assumptions regarding the input load and available resource capacity, which may not be possible in the open Web environment. Feedback control theory, on the other hand, is very robust with respect to assumptions on load and disturbances, but does not explicitly consider timing constraints. It applies a simple trial and error method to correct performance gradually until the desired specifications are met. Guarantees can be made of the eventual convergence of this approach. Very little a priori knowledge is needed regarding the system and its load. Utilizing feedback control theory for Web server QoS control is becoming an increasingly popular research topic. Recent efforts investigate merging feedback control techniques with real-time scheduling theory to produce dynamic resource allocation schemes with provable convergence guarantees and predictable temporal behavior.

A disadvantage of using feedback control theory is that it is primarily a reactive approach. No correction is made to resource allocation until the performance deviates from the desired levels. In an ideal world, a QoS provisioning mechanism should predict likely future performance deviations before they occur and take preventive actions a priori such that no deviations develop. Fortunately, queuing theory offers a predictive framework that allows reasoning about future temporal behavior from stochastic properties of input load. Hence, combining the predictive power of queuing theory with the corrective power of feedback control may produce tighter guarantees on the performance attributes of Web applications. Such a combined approach bas recently been suggested and is becoming an active topic of upcoming research.

Another interesting development in the Web architecture is the rapid emergence of content distribution networks. Currently distributed architectures and protocols for QoS-aware content distribution are still at their infancy. The topic offers interesting challenges in balancing

consistency, content quality, and access timeliness, versus resource constraints in the network such that some notion of global utility is optimized. Such algorithms will undoubtedly receive much attention in the near future.

Although research in the Web QoS arena has been very active, the state of deployment is lagging due to several considerations primarily relating to the difficulty in deploying integrated solutions that combine network QoS, end-system QoS, and QoS-support in the operating system. Fortunately, the networking community has made dramatic strides towards making network support for QoS a reality. This progress further increases the importance of solving the fundamental challenges in building predictable architectures from the end system's performance perspective such that a next generation of Web QoS systems finally becomes a reality. A Particularly good reference for further readings on Web architecture, performance, and QoS issues is the recent book by Krishnamurthy and Rexford (2001).

GLOSSARY

Authentication The act of verifying client identity.

Backbone provider A party that owns the communication fabric of an Internet backbone. Examples of backbone providers include AT&T, Sprint, MCI, and UUNET.

Cache server A network server that acts as a cache of popular Web content and is able to serve it on behalf of the original servers.

Content distribution network A network of server platforms whose sole purpose is efficient content dissemination around the Internet backbone.

Cookies Small text files that servers put on the client's hard drive to save client and session information needed for future access.

Data pull model A data communication model in which the client explicitly asks the server for data each time the data are needed.

Data push model A data communication model in which servers unilaterally push data to the client without being asked. The model is a good optimization when future client requests can be accurately predicted.

Demultiplexing Separating an incoming packet flow into multiple segregated flows. For example, demultiplexing must occur upon the arrival of a packet at a server, in order to queue the packet for the right recipient.

Differentiated services A framework for classifying network traffic and defining different policies for handling each traffic class, such that some classes receive better service than others.

EDF Earliest-deadline-first scheduling policy. As the name suggests, it schedules the task with the earliest deadline first.

HTML Hypertext markup language, a language for defining the content and appearance of Web pages.

HTTP Hypertext transfer protocol. It is the protocol used for Web access in the current Internet. Currently, it has two popular versions, HTTP 1.0 and HTTP 1.1.

IP Internet protocol. It is the glue that connects the computer subnetworks of which the Internet is composed and is responsible for packet addressing and routing between Internet senders and receivers.

IP fragment Part of an IP-layer message after fragmentation.

Kernel The core part of the operating system, typically responsible for scheduling and basic interprocess communication.

Microkernel An operating system architecture where most operating system functions are delegated to user-level processes, keeping the kernel small.

QoS Quality- of service, a term used in quantifying different performance aspects of Web access such as timeliness and throughput.

Packet A unit of data transfer across a network.

Persistent connections Communication abstraction implemented by HTTP 1.1. It allows the same TCP connection to be reused by multiple Web requests to the same server. This is a main departure from the traditional "one request per connection" model of HTTP 1.0.

Proxy server A specialized server that performs an auxiliary Web content management function such as content replication, caching, or transcoding.

Semaphore An operating system construct used for synchronization.

Sockets The main interprocess communication abstraction, originally introduced in UNIX. A socket represents a connection endpoint. The connection is between two processes on the same or different machines.

Threads The smallest schedulable entities in multithreaded operating systems.

TCP Transmission control protocol, the transport protocol used for reliable data communication on the Internet.

Transcoding The process of converting content on the fly from the server's format to a format more suitable to the client, or more appropriate for network load conditions.

Web hosting The business of providing resources (servers, disk space, etc.) to serving customers' Web pages. Typically, Web-hosting companies build large server installations of hundreds of machines for serving the Web sites. These installations are called server farms.

CROSS REFERENCES

See *Circuit, Message, and Packet Switching; HTML/XHTML (HyperText Markup Language/Extensible HyperText Markup Language); Middleware; TCP/IP Suite; Web Hosting.*

REFERENCES

Abdelzaher, T. F., & Lu, C. (2001). Schedulability analysis and utilization bounds for highly scalable real-time services. Paper presented at the IEEE Real-Time Technology and Applications Symposium, Taipei, Taiwan.

Abdelzaher, T. F., Shin, K. G., & Bhatti, N. (2002). Performance guarantees for Web server end-systems: A control-theoretical approach. *IEEE Transactions on Parallel and Distributed Systems, 13*(1), 80–96.

Arlitt, M., Friedrich, R., & Jin, T. (2000). Performance evaluation of Web proxy cache replacement policies. *Performance Evaluation, 39*(1–4), 149–164.

Banga, G., Druschel, P., & Mogul, J.C. (1999). Resource containers: A new facility for resource management in server systems. In *Symposium on Operating Systems Design and Implementation* (pp. 45–58). Berkeley, CA: USENIX.

Barford, P., & Crovella, M. (1998). Generating representative Web workloads for network and server performance revaluation. In *Proceedings of the ACM SIGMETRICS '98 Conference*. New York: Association for Computing Machinery.

Bhatti, N., & Friedrich, R. (1999). Web server support for tiered services. *IEEE Network,* September/October 1999.

Breslau, L., Cao, P., Fan, L., Phillips, G., & Shenker, S. (1999). Web caching and Zipf-like distributions: Evidence and implications. In *Proceedings of the IEEE Infocom '99 Conference* (pp. 126–134). Piscataway, NJ: IEEE.

Chandra, S., Ellis, C. S., & Vahdat, A. (2000). Application-level differentiated multimedia Web services using quality aware transcoding. *IEEE Journal on Selected Areas in Communication, 18*(12), 2544–2465.

Chen, H., & Mohapatra, P. (2002). Session-based overload control in QoS-aware Web servers. In *Proceedings of the IEEE Infocom 2002 Conference* (pp. 516–524). Piscataway, NJ: IEEE.

Cherkasova, L., & Phaal, P. (1999). Session based admission control—A mechanism for improving performance of commercial Web sites. In *Proceedings of the International Workshop on Quality of Service* (pp. 226–235). Piscataway, NJ: IEEE.

Crovella, M., & Bestavros, A. (1997). Self-similarity in World Wide Web traffic: Evidence and possible causes. *IEEE/ACM Transactions on Networking, 5*(6), 835–846.

Kelly, T., Chan, Y., Jamin, S., & Mackie-Mason, J. (1999). Biased replacement policies for Web caches: differential quality-of-service and aggregate user value. Paper presented at the 4th International Web Caching Workshop. San Diego, CA.

Krishnamurthy, B., & Rexford, J. (2001). *Web protocols and practice: HTTP/1.1, networking protocols, caching, and traffic measurement*. Reading, MA: Addison–Wesley.

Lu, Y., Sexana, A., & Abdelzaher, T. (2001). Differentiated caching services: A control-theoretical approach. In *Proceedings of the 21st International Conference on Distributed Computing Systems* (pp. 615–622). Los Alamitos, CA: IEEE.

Menasce, D. A., Almeida, V., Fonseca, R., & Mendes, M. A. (2000). Business-oriented resource management policies for e-commerce servers. *Performance Evaluation, 42*(2–3), 223–239.

Mogul, J. C. (1995). The case for persistent-connection HTTP. *ACM Computer Communications Review, 25*(4), 1995, pp. 299–313.

Web Search Fundamentals

Raymond Wisman, *Indiana University Southeast*

INTRODUCTION

Until the invention of the railroad, a horse's gallop was the limit for speed of overland travel. For finding information, reading the pages of a book was the fastest common means before the invention of the automated search. Search engines are a recent invention, only becoming widely available through the Web, which is itself largely a product of the search engine. Are search engines already changing our source and means of information discovery? The answer is yes, according to some commercial studies that indicate that for many people search engines have already become the main means of seeking information. One report (Lewis, Mobilio, & Associates, 2000) points out that Americans already use search engines 32% of the time when seeking information, more than any other alternative. Much of the information sought is for profit in some way; search engines are the top way consumers find new Web sites online, used by over 73% of those surveyed (Van Boskirk, Li, Parr, & Gerson, 2001).

As with all technologies, history will judge the lasting contribution of the Web and search engines to commerce and society. This chapter's purpose is more modest, being merely to examine the viewpoints of the three partners in Web search: the searcher seeking information, the search engine that locates information, and the Web site holding information.

The reader should keep in mind that the searcher is the only reason for Web sites and search engines to exist and it is in the interest of each to satisfy the searcher. The chapter attempts to expose the search process in a form that is of interest to either the information searcher or the Web site designer. Three main sections divide the chapter. The first section examines the information seeker's viewpoint, considering how to search and measures of a search engine's performance. The second section covers search engine workings and how search engines differ. The third section examines the Web site holding the information, how to be noticed by search engines, how to establish what search engines search, and how to manage or influence search engines to a Web site's benefit.

A broad perspective on the search process is useful before detailed examination of these points of view and their interactions. Figure 1 illustrates the characteristic interaction between these three partners in information search and delivery. The first step occurs when a Web site creates pages with information and sends the main Web site page location to a search engine. Next, in the typical search engine architecture (Arasu, Cho, Garcia-Molina, Paepcke, & Raghavan, 2001), one component called a spider (or crawler, robot, etc.) visits the Web site to retrieve pages linked from the main page much as a person using a browser would follow links to new pages. The spider follows links, indexing meaningful words of the retrieved pages which are stored along with the page's Web site location for later searches and retrievals. The searcher can then send queries to the search engine and receive a list of matching Web site page locations, from which the searcher selects and retrieves pages directly from the Web site. Closing this information space between the searcher and Web site is the primary purpose of the search engine. The following sections attempt to provide insight into the full search process so that the information searcher can find better information more easily and the Web site designer can better attract searchers to the information.

HOW TO SEARCH—THE SEARCHER'S VIEW

Why search the Web? A simple answer is that the Web is too large and unorganized to find much useful information. In 1999 public sites contained about 800 million pages on about 3 million servers (Lawrence & Giles, 1999). The original Web design purpose was simply to interconnect bits and pieces of scattered information with no plan to find information other than by manually moving from one piece of information to another by following connecting links. The search engine merely automates the process and moves much faster from one piece to the next than users do, while collecting information along the way for later retrieval. Web site usability studies recognize search

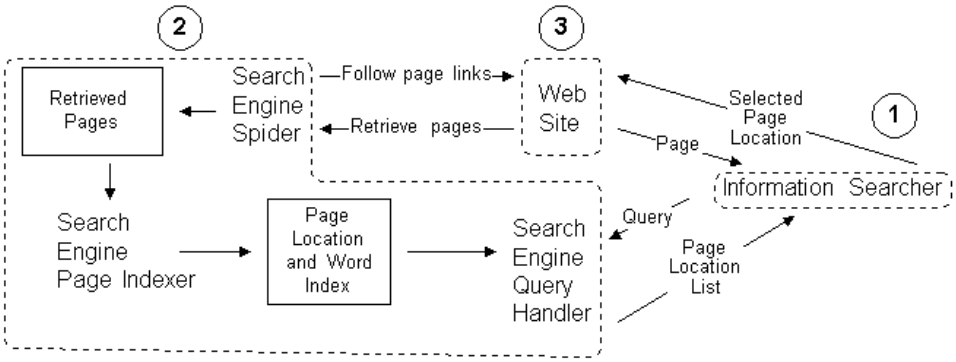


Figure 1: An overview of a search engine and its relation to Web sites and information searchers. The basic steps to Web search are as follows: a search engine spider visits a Web site to retrieve and index pages for later searches, the searcher sends queries to the search engine and receives a list of page locations, the searcher selects and retrieves pages from the Web site.

as one of the most important means for finding information (Nielsen, 1997). More than half of all Web site visitors navigate predominately by search and use search as their first choice to find specific information. Many usability experts argue that search is so important it should be available from every single page on a Web site. Part of the rationale is that once visitors navigating connection links get lost they will need search and should not have to search for a search page.

Types of Search

Search often fails to some degree. To identify good search strategies it is worthwhile to consider the different types of search and how search engines match different strategies. Information retrieval experts (Rosenfield & Morville, 1998) recognize four ways in which people search:

Known-Item Search: A search engine is not always necessary. Searching a brokerage site for the latest stock quote is one example. The information sought is unique, is easily recognized when found, and often exists at known Web sites. Searching, if any, is short and direct and the answer immediately recognized. Search is unneeded if the information location remains fixed, even as the information changes. The search can be restricted to specific sites, reducing search time and improving the information quality.

Existence Search: Consider the difficulty in searching for something that may not exist. Suppose that someone who needed to organize phone number lists, notes, and appointments had never heard of a personal digital assistant (PDA). The person would know the needed function but not where to look or that a PDA even existed. Search can meander as users discover and digest information; even recognizing the answer when found may be difficult. Web search engines are a natural choice for determining whether information exists because they attempt to find all available information on the Web. Entering a few well-chosen query words to a search engine will search millions of pages and can produce thousands of pieces of information. One gap in existence searching is that search engines cover only a fraction of the Web; the more obscure information can exist on parts of the Web never visited by the search engine.

Exploratory Search: This is the case when the user knows that the subject such as a PDA exists but does not know exactly what a PDA can do and need to learn more. Uncovering and comprehending the information may be long and circuitous; recognizing the answer may not be immediately possible due to the lack of subject knowledge. The most productive search strategy is browsing information organized into subject lists with main subject areas and a breakdown into finer categories of related subjects.

Comprehensive Search: This type of search is best suited when users know the search subject, can recognize the answer when found, and need all available information on the subject. For someone with a general knowledge of PDAs, buying a PDA requires comprehensive information on several brands, models, and suppliers to make an informed decision. Web search engines work well for comprehensive searching of the entire Web but for broad subjects such as PDAs can produce an overwhelming volume of information.

Information Sources

Most successful searches fail to some degree at the first attempt. Other than the known-item search, most searches evolve to use several search strategies as one attempts to iteratively improve the quality of information retrieved. Because Web search sites accommodate one or more of the search strategies we will focus on the type of sites best suited to users' search needs, the most common ways to search, and techniques for successful searches.

Though the term search engine often applies to any computerized information source, there are at least three different types of Web information sources: automated search engines, human-organized lists, and portals. Most information sites are a mixture of some or all three types; the following characterizes each.

Lists

Catalogs, directories, and lists are human-organized subject indexes of Web information. A list arranges subjects into a hierarchy that allow exploration from a general topic such as "education" down to the more specific subtopic of "graduation poems." On most lists, humans index the subjects providing control and intelligence to the subject hierarchy organization, arguably yielding less but higher quality information than current automated cataloging systems. The hierarchical approach works well for exploratory search when one is familiar with the search subject and has a good sense how the subject fits within the hierarchies of a larger subject.

Finding a subject in a hierarchical list can be difficult when the searcher's idea of subject organization differs from that of the person creating the list. For those cases, a keyword query can search the subject list as a way to pierce an opaque subject hierarchy; a query such as "graduation poems" will generally locate the same information as manually moving through the "education" subject hierarchy. If the query fails, most list services automatically send the query to a Web search engine; for example, the Yahoo list currently collaborates with the Google search engine for Web searches of failed subject queries.

Search Engine

A search engine matches query words with words on Web pages and lists the pages containing the matching words. Entering the query word "zucchini" returns a list of pages containing the word "zucchini." In practice, Web search engines examine vast numbers of pages while automatically calculating page importance to rank each page from highest to lowest to reduce quantity and improve information quality. Accurately determining page rank based on a few query words is challenging and sometimes produces completely irrelevant results. Entering the single word "zucchini" will find any page mentioning "zucchini," from gardening to cooking to diet; which "zucchini" page is important is ultimately in the mind of the searcher. Determining how to rank one document against thousands of others is a key point of competition between search engines and often the identical search on any two will produce different results, searchers may need to consult several to have confidence in the results. Metasearch engines automate searching on multiple sites by sending the search query to a number of search engines and creating a fusion of the results. Overall, search engines are best suited for existence and comprehensive searches where the quantity of information is of prime importance.

Portal

A portal attempts to make using the Web easier. Portals attract a broad range of visitors by serving as information department stores with a variety of services, such as e-mail, forums, search engines, subject lists, and online shopping malls. Portals are also distinguished from pure information sites by the degree of personalization possible for information sources by combining search to produce personalized services for local weather, stock portfolio tracking, calendars, and search categories. Portals often integrate external search engines with internal subject directories and other information tools to improve quality and a seamless interface to a set of multiple services. Portal tool integration and ease of use has transformed the Web into a household appliance, one with obvious benefits and easily understood and used by the broader population (Hock, 2001). Since 1998, portals have been a major force driving the rapid growth of the Web.

America Online (AOL) was among the first portals and has one of the most extensive lists of services and degree of personalization. Today, most major search engine and list sites now offer many of the same services to attract and keep revenue-generating visitors. Other portals target visitors with narrower information and service needs; for example, a financial service portal would provide a range of personalized financial services along with focused financial information. Even as competition blurs the differences between portals, lists, and search engine sites, information in some form remains the primary lure to attract visitors to each.

How to Use Search Engines

What is the best search strategy? For a clearly defined subject area, use a site that maintains subject lists. On Yahoo, one can find "weight loss" carefully placed in the "Health > Weight Issues > Weight Loss" subject hierarchy. The "Weight Loss" category then has seven subcategories ranging from liposuction to diets and a list of some 30 other Web sites devoted to weight loss. Human thought in placing each subject in the organization is obvious as each subject relates in some way to others nearby. In comparison, a search engine can find over 1 million pages with the words "weight loss." Although the search engine calculates the rank of each page, there is no guarantee that one page relates to another except through the common coincidence of the words "weight" and "loss."

When are search engines the better choice over lists? With no clear definition of the subject area, subject lists are of little value. Consider locating genealogy information for a family. Unless the family is already famous, finding a subject list devoted to it is doubtful, but entering the family name on a search engine would likely locate dozens of individual pages and Web sites. A search engine is better suited for finding comprehensive information that does not fit into a clearly defined category.

How to Search

One of the main problems with search engines is they are much better at finding a high quantity than a high quality of information. To understand how this affects search strategies, consider that search engines operate by keyword searching, searching for pages containing one or more of the words entered as a query. To research diet plans, a searcher might enter the query "weight loss" and the search engine might return several million page references, each containing the word "weight" or "loss." Can search be limited to pages that are useful? The answer is yes, but because search engine strategies purposely differ, there is not a single protocol for defining a search. There are, however, some common approaches to improving search success.

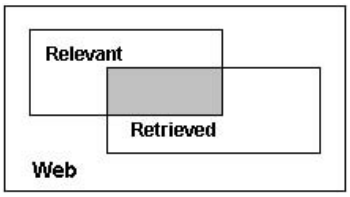


Figure 2: Retrieved versus relevant references when searching the Web. The intersection of retrieved and relevant pages will likely be relatively small in part because only a fraction of the Web is indexed.

Several basic definitions common in the search literature, "relevancy," "recall," and "precision," are also of value to searchers. Relevancy measures the usefulness of the references retrieved and is highly subjective. Figure 2 illustrates three useful points: (1) that relevant references are generally a subset of all the references available; (2) that retrieved information generally includes some irrelevant references; and (3) that the intersection of retrieved and relevant references will often exclude references that are relevant (Belew, 2000). The following definitions of *recall* and *precision* nicely complement the intuition illustrated in Figure 2:

$$Recall = \frac{|\text{Retrieved} \cap \text{Relevant}|}{|\text{Relevant}|}$$

$$Precision = \frac{|\text{Retrieved} \cap \text{Relevant}|}{|\text{Retrieved}|}.$$

Recall is the percentage of relevant references retrieved. A recall of 50% means that only one-half of the relevant references were retrieved; retrieving every possible reference yields 100% recall but often includes many references that are irrelevant. Precision is the percentage of retrieved references that are relevant. Of the references retrieved, a precision of 50% means that only one half of the references were relevant; returning only one relevant reference yields 100% precision but possibly very poor recall. A "weight loss" query can yield perfect recall by retrieving every Web page in existence but will likely produce many that are not relevant, so that precision is poor. Simultaneously achieving perfect recall and precision is nearly impossible even when carefully designed and constructed sets of known information are searched. Given the scope, size, disorganization, and diversity of information sources available on the Web, merely retrieving relevant information is daunting.

Search engines naturally exhibit great recall when finding millions of pages for a broad query, but narrowing the query can greatly improve precision by eliminating many pages of questionable relevancy. Fine control of a search can yield higher quality references, is relatively easy to use, and is common to most search engines. Table 1 lists common search controls for popular engines with details and examples of search refinement provided in the following discussion (Sullivan, 2001).

Adding Search Terms: The simplest and often most effective method of improving the precision of search results is adding search terms with a more precise meaning. The query "weight loss" produces millions of pages using the Teoma search engine, each page containing the word "weight," "loss," or both. Adding terms for "vegetarian weight loss diet" produces merely 10,000 but more focused references. Each page found still contains at least one or more of the query words. Because search engines generally rank pages higher that match more query terms, pages with fewer discriminating terms are effectively ignored.

Searching for Phrases: Phrases of specific word groupings can yield more precise searches than the same words when matched independently. Searching on the

Table 1 Common Search Controls for Popular Engines

Query Match	How	Example	Popular Search Engine Support
Any	Automatic	vegetarian diet	Teoma, Yahoo, Inktomi
	OR	vegetarian OR diet	AltaVista, Yahoo, MSN, Google
All	Automatic	vegetarian diet	AltaVista, Google
	AND	vegetarian AND diet	AltaVista, Yahoo, MSN
Include	+	+vegetarian diet	AltaVista, Yahoo, MSN, Google, Inktomi
Exclude	−	vegetarian–diet	AltaVista, Yahoo, MSN, Google, Inktomi
	AND NOT	vegetarian AND NOT diet	AltaVista, MSN, Inktomi
Phrase	" "	"vegetarian diet"	Google, Yahoo, Inktomi, Teoma
Wildcard	*	veget* diet	Altavista, Yahoo, MSN
Proximity	NEAR	vegetable NEAR diet	AltaVista
Title	title:	title: "vegetarian diet"	Altavista, Yahoo, Inktomi
	intitle:	intitle: "vegetarian diet"	Google
Site Only	site:	site:food.com vegetarian	Google, Inktomi
	host:	host:food.com vegetarian	AltaVista
URL Only	url:	url:food.com	AltaVista, Yahoo, Inktomi
	allinurl:	allinurl:food.com/diet/	Google
All References	link:	link:www.food.com	AltaVista

phrases "vegetarian diet" "weight loss" produces only about 3,000 page references, of which the highest ranked pages contain both of the two phrases.

Inclusion and Exclusion: Query words prefixed with a "+" are words that must be included or found on a Web page, whereas words prefixed with a "−" must be excluded or not occur on the page. The query "+diet −vegetarian" then means that "diet" must be found on the page and "vegetarian" must not be found. The Boolean operators AND, NOT, AND NOT are often used alternatives to the "+" and "−" operators. The query "diet AND NOT vegetarian" is normally equivalent to our earlier query. The OR operator matches either term; "weight OR mass" matches "weight" interchangeably with "mass," much to the annoyance of physics teachers.

Proximity: Limiting matches to only words or phrases occurring within a close proximity to another assumes that nearness implies some relation between the words. The query "vegetarian NEAR diet" would find pages having "vegetarian" within a few words of "diet" but exclude pages having "diet" distanced from "vegetarian" by more than the proximity word limit.

Wildcard: An "*" at the end of a word or partial word expands the range and number of matching words. For example, the query "veget*" finds pages that include "vegetarian," "vegetable," and any other words starting with "veget."

Field: Limiting the search to a designated field of the Web page, such as considering only pages available on a specified Web site, can greatly narrow the search focus. As an example, title: "vegetarian diet" requires "vegetarian diet" to be part of the page title field, whereas "site:food.com" limits the search to the "food.com" Web site pages. The link control lists all pages with a reference link to a specified site; the query "link:food.com" will list all pages that link to pages on the "food.com" Web site.

Combined Controls: Combining multiple search controls yields a more discriminating search. The query "vegetarian diet domain:au" limits search for vegetarian diets to only Australian sites; Australian-based sites end with "au" in the domain name.

Search Engine Performance

A single search engine never performs the best in all cases; a search that fails on one can succeed on another. Search engines compete based on the scale and strategy of search; finding the best pages for the searcher is not only a point of technical distinction but also a competitive advantage. Using many different search engines for a single search is one strategy for improving the probability of finding relevant information, which is basically what a metasearch engine does by automatically submitting a query to multiple search engines and presenting the fusion of the highest ranked results from each.

Performance Measures

The three important performance measures are recall, precision, and ranking. As discussed earlier, recall is defined as the percentage of relevant pages found and precision as the percentage of pages found that are relevant. A search engine calculates rank to measure relative page relevancy, using individual page rank to order the pages from less to highly relevant. As a way to gauge individual search engine strengths and weaknesses, the following describes searches that provide observations of these three measures.

Recall: Searching for something one knows is on a Web site gives a rough estimate of recall. If a user's name appears in Web pages on his or her site, searching on the name should return results from the site and perhaps other sites as well. Restricting the search to the user's specific site, if recall is 100%, should return all pages on the site containing the user's name. For testing total recall, the query "vegetarian site:www.food.com" should return all pages with the word "vegetarian" from the site "www.food.com." Most search engines follow only a limited number of page links on a single site, stopping at some maximum number of links deep. Searching for pages several links deep from the site main page then measures the search engine's recall ability. If one-half of the relevant pages on a site are found, recall is 50% and indicates that the search engine spider stopped after following some arbitrary number of links from one page to the next.

Precision: Precision is difficult to mechanically quantify as it measures the number of relevant pages among those found and relevancy is by nature subjective. Because search engines generally list only pages that contain matching query words, precision is always arguably high. However, when users find thousands of pages and only a few are relevant to their needs, the challenge is to focus the search on more relevant and generally fewer pages. The searcher can influence the search precision by including or excluding query words, limiting the search to known sites, and using other controls discussed earlier. Searching for a known page can test the degree of search control and precision afforded by a search engine. Most search engines provide sufficient control to limit the results to a single page. For example, on some search engines the query "vegetarian diet" would find all pages with "vegetarian diet" anywhere in the page whereas the query title: "vegetarian diet" would find only pages with that phrase in the page title. The more accurate the search controls, the better the search precision.

Rank: Rank reflects the calculated relevancy of a document. One key factor determining rank is the number of query words matched in a page; generally the more words matched the higher the rank. Matching rare words also increases the rank of the page.

Search engines can also employ the structure of the Web to improve relevancy based on links to and from other pages to augment the ranking of pages. Pages with many links from other pages generally rank higher than pages of equal similarity to the query but with fewer links.

Page popularity and importance are two common ranking measures based on link structure. Popularity assigns a higher rank to pages having more references

from other pages, under the assumption that a frequently referenced page is more relevant than a page with fewer references. However, a popular page is not automatically important. Page importance further assumes that pages or sites that mutually refer to one another form groups of importance or authority on a subject. The more references to a page from others in the group, the higher the importance and consequent rank of the page. The resulting groups are somewhat analogous to the subjects of a list service with the exception that humans organize subject lists and the interconnecting references organize a group. Automatic grouping can produce more focused and higher quality references but eschew the hierarchical organization that supports browsing of subjects by the searcher.

Which ranking approach produces the best result depends upon the user's search needs. Comprehensive search is the natural outcome of search based on word match ranking alone but yields no organization of the results. Existence and exploratory search can benefit from the reference-based ranking methods of popularity and importance. Popularity ranking anticipates that the information that many others reference represents common knowledge of a subject. Importance attempts to refine popularity ranking by organizing references into supporting groups. Grouping together documents that have common references will generally provide more homogenous results and is best suited for exploratory search where the subject is recognized.

HOW SEARCH WORKS—VIEWS FROM THE SEARCH ENGINE

Although human-organized lists and automated searches are often lumped together as search engines, how each gathers and represents information is quite different.

Human-Organized Lists

Human-organized lists depend upon editors to first review Web site pages and organize the collected information into subject hierarchies. A searcher can then manually browse a subject hierarchy list to find information. Lists more accurately reflect true page content than automated searches because human editors do not review the parts of a page not normally seen by a reader and can ignore gratuitous or repetitive words used to attract a search engine. One weakness of lists is the effort required of the editors to perform the review. Once a site review is completed, significant changes to the site content are unlikely to prompt another review, although most lists do accept site resubmissions.

Search Engines

Automated Web search engines have two main tasks: placing keywords from each page into an index and answering search queries from the index. Web-based search engines are only recent versions of traditional full-text systems, such as the SMART and SIRE systems (Salton & McGill, 1983). Rather than employing human editors to extract and index subject information from pages, such systems perform automatic indexing of complete pages into a database, creating lists of words and the page locations that contain each word. Answering queries from the index compares query words to page words, retrieving those pages that have a high calculated similarity to the query.

Indexing

The automatic indexing of a page first removes most common or stop words such as "a," "the," and "it", because almost all pages contain those words and they are nearly useless in discriminating between pages. Each remaining word then has a weight factor calculated based in part on word frequency. Weight measures also may give greater or lesser importance to words in different parts of the page; for example, title words generally receive more weight than regular text words. A word that occurs in few pages also has a higher weight, based on the proportion of times that the word occurs in a single page relative to all other pages. A word that occurs on only one page would have relatively high weight whereas a word occurring on all pages would have low weight, the rationale being that a rare word is more useful in finding a page than a word common to many pages. The same rationale applies to the removal of stop words.

Conceptually, the resulting index contains all words from all pages, excluding stop words. Included with each word is a list of all the pages that contain the word. For each page in the list, there is stored the frequency with which the word occurs in the page for weight calculation and the location of the page for retrieval.

Web search engines differ significantly from traditional search engines in that pages are scattered across many thousands of Web sites and each page can have connections to many other pages. To find pages and build the index, the indexing program must then visit Web sites as one would with a browser, starting at a page, visiting connected pages, and indexing each page as it goes. In the jargon of the Web, a spider, robot, or Web-crawler is the program that visits or crawls connected pages, indexing selected parts from the visited pages. The resulting index contains the word lists of traditional text searching and the unique location of the page for retrieval. For each page indexed, the connecting references to other Web pages are included when the popularity and importance of pages are to be calculated.

Retrieval

The traditional retrieval process produces a ranking of all pages that contain one or more query words entered by the searcher. The retrieval operation consists of converting the query words to the same representation as the pages, calculating a similarity measure between the query and each page in the collection, and retrieving pages from the collection ranked in order of high to low similarity.

A common similarity measure is the cosine of the angle between the words of the query and the words from individual pages. Determining the cosine similarity measure requires representing the words of the query and pages as vectors in a multidimensional word space where each axis corresponds to a different word drawn from the indexed pages. Roughly, if a page matches the query, the angle between the query and the page vector is zero; having

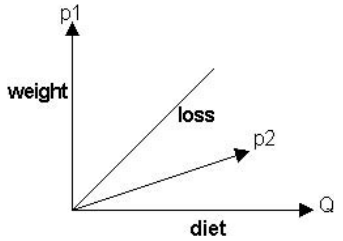


Figure 3: Two pages indexed on the words "weight," "loss," and "diet," where page p1 contains the word "weight" and p2 contains the words "loss" and "diet." Similarity is measured as the cosine of the angle between a page and query vectors; the smaller angle corresponds to greater similarity. A query of "diet" would have some similarity to page p2 but no similarity to the "weight" page p1.

fewer words in common produce a greater angle or smaller similarity. Figure 3 illustrates the indexing of pages p1 and p2 containing the words "weight" and "diet loss" respectively, where each axis corresponds to one of the three words. Representing the pages and query as vectors in three-dimensional space, the query vector for "diet" has a smaller angle with respect to the "diet loss" page p2 than with respect to the "weight" page p1. Intuitively, the "diet" query is somewhat similar to the "diet loss" page and not similar to the "weight" page. Based strictly on similarity to the query "diet," the search engine would return a high ranking for the p2 page and a low ranking for the p1 page.

Web search engines can exploit the special features of Web pages to further improve retrieval quality beyond that possible using query and page words alone. Specific elements of a page, such as title words, can augment strict similarity measure ranking. When a page is retrieved, the engine generally returns the calculated ranking, designated parts of the page such as the title and a description of the page content, some text surrounding the word found, and most importantly, the location of the page to retrieve.

Page popularity and importance measures utilize the natural links between Web pages to refine rank calculated on word matches alone. Popularity can be defined to assign a higher rank to pages having more references from other pages, under the assumption that a frequently referenced page is more relevant than a page with fewer references. Figure 4 illustrates the popularity of page A to be greater than that of either B or C, as more links or references are to A than to either B or C.

However, a popular page is not automatically important. Importance of a page can be defined in terms of the number of links to the page and the importance of the linking sites. For example, a link from an Internal Revenue Service page to a page on taxes is intuitively more important and adds greater rank than a link from a random individual to the same page. This global ranking scheme is the basis of PageRank (Page, Brin, Motwani, & Winograd, 1998), used by the search engine Google.

Although a discussion of the full PageRank algorithm is beyond the scope of this chapter, the basic concepts (Arasu et al., 2001) are relatively clear when limited to a Web where every page can be reached by every other page. For that assumption, let 1, 2, 3, . . . , m be pages on the Web; let N(i) be the number of links from page i; let B(i) be the set of pages linking to page i. Then r(i), the simple PageRank of page i, is

$$r(i) = \sum_{j \in B(i)} \frac{r(j)}{N(j)}.$$

Note that dividing by $N(j)$, the number of pages linked from a page, reduces the raking contribution of a page as the number of pages linked to increases. This fits one's intuition that a page with links to many, possibly loosely related pages should not contribute as much rank as a page that links to only a few, likely more closely related pages.

Authority ranking assumes that a page referring to another confers some measure of authority to the page. Authority and hub ranking utilize page links to identify authoritative pages on a subject and hub pages that link to many of these authoritative pages. Mutually referring pages then form authoritative groups on a subject (Kleinberg, 1999), where the more links to a page from others within the authority group, the higher consequent rank of the page. Figure 5 illustrates a simplistic method of forming groups through mutual B and C links and common BC links to A. Page D marginally contributes to the authority of A but is not included in the ABC group as no other group member references D.

Hub pages are defined to have extensive links to authority pages and serve to bring together authorities on a common subject to form a mutually self-referential community. In Figure 5, page D acts as a hub that connects two separate groups into a single community. The resulting community is somewhat analogous to a single subject within a list service except that humans organize subject lists and the interconnecting links organize a subject community. Automatic grouping can produce more focused and higher quality references than strict similarity ranking but eschew the hierarchical organization that requires subject knowledge on the part of the searcher.

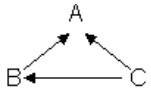


Figure 4: The number of links to a page measures its popularity. Two links confer on page A the greatest popularity, B the next, and C the least popularity. The ranking based on popularity would be ABC.

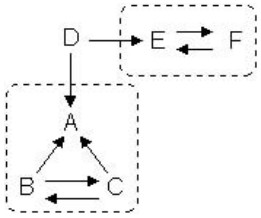


Figure 5: Pages ABC and EF are two separate authority groups connected by hub page D to form a common subject authority.

Automated Search Engines
Automated Web search engines have two main tasks; one of indexing the Web information, the second of answering search queries from the index. First, an indexing program visits a website much as you would with a browser, normally starting at the default homepage, visiting connected pages and indexing the site information (see Figure 6).

Figure 6: How the HTML of Figure 7 would appear in a browser.

Which ranking approach produces the best result depends upon the user's search needs. Comprehensive search is the natural outcome of search based on word match ranking alone but yields no organization of the results. Existence and exploratory search can benefit from the reference-based ranking methods of popularity and importance. Popularity ranking anticipates that the information that many others reference represents common knowledge of a subject. Importance attempts to refine popularity ranking by organizing references into supporting groups. Grouping together documents that have common references will generally provide more homogenous results and is best suited for exploratory search where the subject is recognized.

What Search Engines Search

Web pages can be richer sources of search information than traditional documents such as books and journals because of the natural connections formed to related pages and the characteristics of the hypertext markup language (HTML) used for writing Web pages. Web search engines seek to improve upon traditional retrieval systems by extracting added information from the title, description, and keyword HTML tags and by analyzing the connecting links to and from a page.

Recognizing the parts of a Web page that attract the attention of indexing spiders is critical to Web site designers attempting to raise the visibility of the Web site. Ideally, a Web site designer could give instructions to visiting spiders on precisely how best to index the page to produce high quality search results. Unfortunately, self-promoting Web sites generally have a history of hijacking spider indexing rules for their own benefit. In response to blatant self-promotion, few spiders observe a strict protocol as to which page to index or which parts of the page are considered important. However, most spiders do observe the following common guidelines (Sonnenreich & Macinta, 1998):

Content

The result of search is the page content that the searcher sees and reads. The readable text, as displayed by a browser in Figure 6, provides the bulk of the words indexed by the spider. As noted, stop words are worthless in distinguishing one page from another and are ignored. Less common words increase the page rank but are valuable only if a searcher uses that word in a query. Using many different words in a page improves search breadth but the words must be obvious for a searcher to use in a query. Including important keywords in the title, increasing the frequency of a keyword the text, and placing keywords near the beginning of the page content can improve page rank on most search engines. Be aware that repeating a keyword multiple times in the title may gain a higher ranking but many search engines ban blatantly bogus attempts at manipulation and may reject the page or site entirely. The challenge to the page writer is to find the right keywords rare enough to stand out, descriptive of the content subject and that are familiar to the searcher. Bear in mind that most indexing spiders only examine the first few hundred words of content so it is important to provide descriptive keywords early in the content text.

Tags

HTML tags are not generally visible to the reader but do contain information important to the spider. Along with content keywords, spiders also extract the page location and may examine HTML tags when indexing a page. The Web site designer can influence the page rank and provide more descriptive results to the searcher through the tags. Figure 7 gives the source for a HTML page to illustrate the page content and use of the following tags.

Keyword: The HTML keyword meta tag contains human-defined keywords to augment the automated indexing of the page content. One use of the tag is to provide alternative words or phases for those in the content, for example, using "PDA" in the content and "personal digital assistant" as keywords. Unfortunately, promoters have so often abused the keyword tag that Web search engines generally ignore it. When search is limited to a trustworthy site, such as a university Web site, keywords can be valuable to the designer and searcher.

Description: The description meta tag provides a short content summary for display when the search engine retrieves the page. Figure 8 illustrates how a search engine would display the description tag with other page information.

Title: Indexing the title tag independently allows explicit searches on the title; the search engine can also display the title as part of the page information, as in Figure 8. As previously mentioned, keywords placed in the title can also improve page rank.

Heading: The large print of headings catches the attention of the reader and also is important to an indexing spider. The influence of headings on rank generally follows the scale of the heading number, so that weight of the words of a level 1 heading is greater than the weight of the words of a level 2 heading.

Links: The spider follows link connections to other documents through the attribute and hypertext reference tag; for example, "<a href = 'Figure6.html'>" directs the spider to follow the link to the index page "Figure6.html." The popularity and importance methods would generally rank a page with many links from other pages

```
<html>
<head>
  <meta name="description" content="Human lists and automated search engines.">
  <meta name="keywords" content="search engine, indexing">
  <title>How Search Engines Work</title>
</head>
<body>

<h1>Automated Search Engines</h1>

Automated Web search engines have two main tasks; one of indexing the Web
information, the second of answering search queries from the index. First, an indexing
program visits a website much as you would with a browser, normally starting at the
default homepage, visiting connected pages and indexing the site information (<a
href="Figure6.html">see Figure 6</a>).
</body>
</html>
```

Figure 7: An HTML page contains visible parts displayed by the browser and hidden parts that can help spiders index the page more accurately, provide descriptive information to the searcher, and link to other HTML pages.

relatively high. Influencing other sites to link to a site's pages is not easy, depending upon good content to attract recognition. Cross-listing agreements with other sites to link to a site is one technique that can provide a quicker start to that recognition.

What Search Engines Ignore

Can designers make their Web sites invisible? Are all spiders the same? Not every word on a Web page will lead the searcher to the page. Most spiders purposely ignore or cannot see large parts of a page that human readers might see and use; carelessness in page design can force both spiders and human searchers away. The most common spider problems and solutions are the following:

Frames: The purpose of frames is to visibly divide a browser screen into several parts, but unfortunately frames can stop an indexing spider and create confusion for visitors arriving from a search engine. At least three separate pages are needed for frames: a hidden frameset page that defines the frames and links visible pages into the frames, a second page for visible content, and a third that is often for navigation. A spider normally arrives at the hidden frameset page but must understand how to handle frames in order to follow links to the other, visible pages. Spiders that do not understand frames simply stop and never index further. For those spiders and browsers that do not understand frames, the remaining site pages may be unreachable. Because frames cause some spiders problems, the obvious solution is to avoid the use of frames entirely. However, when the Web site designer is forced to use frames, including the "noframes" tag exposes alternative text for navigation and content to frame-ignorant spiders and browsers. The "noframes" tag designates text to be displayed in place of the framed pages, effectively duplicating the page content and design effort. Spiders that understand frames create a different problem. Visitors can now arrive at a content page directly from the search engine rather than through the main frameset page as the Web site designer intended. Without the main frameset page there is no navigation page either; visitors can become wedged on a dead-end page and, without any navigation, forced to leave the site. One solution is to place a link to the main frameset or site home page in every navigation and content page to help keep the visitor on the site. Of course, another solution is to avoid frames altogether.

```
How Search Engines Work
Human lists and automated search engines.
http://www.insearchof.org/how.htm
```

Figure 8: An example of how a search engine might respond to a query. The word "indexing" is part of the keyword meta tag and embedded in the text content. The title is "How Search Engines Work" and the description meta tag is "Human lists and automated search engines." The document URL "www.insearchof.org/how.htm" and the title provide links to the complete document.

Scripts: Most spiders ignore script programs written in JavaScript or other scripting languages; others simply index the script program text. Spiders that index the script may also index only the first few hundred words of the document and possibly never reach the content. Place important content and keywords before scripts and, for pages that are mostly scripts, include title, keyword, and description tags.

Java Applets and Plug-Ins: To a spider, a Java applet, plug-in, or other browser-executed program is invisible. For indexing purposes, include descriptive content and tags within the page that contains the program.

Server-Generated Pages: Spiders may ignore any unusual link references, such as ones that do not end in "HTM" or "HTML." For example, a spider will follow the connecting link "<a href = 'mainpage.HTML'>" but may not follow the link to the Web server program of "<a href = 'mainpage.ASP'>" due to the ending "ASP." Generating the Web site main page with a server program could mean that some spiders ignore the complete site. One solution is to provide some

HTML pages for the spider to index and eventually guide the searcher to program-generated pages.

Forms: Collecting visitor information is one of the most important functions of many Web sites, but spiders do not know how to fill out forms; a site login form automatically stops a spider. Spiders that do index the content and links of the page containing the form create potential problems by leading visitors directly to the form page from a search engine rather than through pages intended to precede the form. An example would be an airline reservation system with forms for itinerary and payment. Visitors arriving directly at the payment form would obviously have problems; consider adding links back to a starting page.

Robot Exclusion: Forms represent one good reason to exclude spiders from indexing certain pages. Two standards exist that instruct well-behaved spiders on excluding specified pages. The recognized standard is the "robots.txt" file that lists acceptable and unacceptable directories, pages, and robots (i.e., spiders). A single robot.txt file exists for the entire Web site, which only the site administrator can access, creating a maintenance bottleneck as multiple designers make changes to the site. A better but less accepted solution defines a special "robots" meta tag to specify how to index each individual page. Options are that every spider or no spider should index the page, the page should be indexed or not, and the page links should be followed or not.

Images: Spiders may index the image location, image title, and alternate text but that is probably all. For searchers to find the image, include some additional information about the image in the page content and the HTML tags.

Deeply Linked Pages: Most spiders completely index only small sites, generally indexing only limited pages on each site. Spiders limit indexing to several connecting links deep, ignoring pages linked beyond that depth. As a rule, keep pages important in attracting visitors linked directly from the site home page.

Reorganization and Broken Links: Resist the urge to reorganize the site. Until all the spiders come again, the new greeting for visitors arriving from search engines to pages that have been renamed or otherwise permanently hidden may be "404 Not Found." Adding new pages to the site is fine; just leave existing page locations and names alone.

Metasearch

Individual search engines produce results biased in ways that are unpredictable and invisible to a searcher. Sending the same query to several search engines will obtain widely different results that clearly illustrate the bias. One study on search engine bias (Mowshowitz & Kawaguchi, 2002) demonstrated that querying nine popular search engines for information on "home refrigerators" produced 14 different brand names in the cumulative top 50 results of each engine. Reporting of the brands was unpredictable and uneven across search engines consulted; several brands were found by only a single search engine and no search engine found a majority of the 14 brands. And obviously there was no clue as to the brands not found at all.

Given that individual search engines index only a small fraction of the Web and the degree of index overlap among a group of search engines may be small, it makes sense to consult multiple search engines. Metasearch engines automate multiple searches by sending the query to several standard search engines and organizing the fusion of the results uniformly. Unfortunately, metasearch engines cannot broadcast a query to all other search engines but attempt to minimize the use of limited resources such as network bandwidth while maximizing information quality. Although increasing the number of information sources will generally improve recall, it is also likely that precision will suffer correspondingly. Balancing these conflicting goals is the key challenge in designing a metasearch engine.

The essential architecture of a metasearch engine (Dreilinger & Howe, 1997) consists of a dispatch mechanism to determine which search engines receive a specific query; interface agents to contact and adapt the query and result formats of each search engine; and a display mechanism that creates and displays a uniform ranking of results after removing duplicates. By depending upon the direct results returned from regular search engines, metasearch engines cannot expand or improve upon the information sources. However, to a searcher, metasearch represents an obvious improvement over a single search engine by the simple increase in the number of search engines consulted and the corresponding increase in the fraction of the Web examined.

HOW TO BE SEARCHED—VIEWS FROM THE WEB SITE

The purpose of building a Web site is to attract visitors; information is the lure. Visitors most often find a new site via a search engine, so building an easily found and searched Web site is critical. A study of search success (Users Don't Learn to Search Better, 2001) illustrates the challenges of designing a Web site for search. After watching 30 searchers search different sites for content that was on the sites, the study concluded: "The more times the users searched, the less likely they were to find what they wanted." Single searches found the content 55% of the time, those searching twice found the content only 38% of the time, and those searching more than twice never found the content. Nearly 23% of the searchers received a "no results" response on their first searches, causing most to give up immediately. For those who continued to search, results only grew worse. Further compounding search problems is the prevalence of invalid links to pages that are no longer accessible; one study (Lawrence et al., 2000) gives the percentage of invalid links ranging from 23% of 1999 pages to 53% of 1993 pages. The collective message seems clear: design the site and pages for search and continually test that search works.

Designing a Web site for search is the subject of this section; the details of page search were covered in the previous section. This section divides search of a Web site into two main parts: search that includes the Web site as

Table 2 Server Access Log Fields in Common Logfile Format

Access Log Field	Example
Client IP address	24.10.2.3
Client identity-unreliable	–
Authenticated client userid	–
Time request completed	[01/May/2002:17:57:03 -0400]
Client request line	"GET/mainpage.htm HTTP/1.1"
Server status code	404
Referring site	"http://www.food.com/search.html"
Client browser	"Mozilla/4.08 [en] (WinNT; I;Nav)"

part of the Web and search that is restricted to a single Web site.

Web Site Discovery

How is a Web site discovered by a search engine? Given that any one search engine indexes only a small fraction of the Web (Lawrence & Giles, 1999), the answer is of critical importance to the Web site designer hoping to attract visitors. Most search engines accept free submissions for indexing all or part of the Web site and paid submission to multiple search engines is available through service companies. Links from other sites will also widen visibility and speed the discovery of a Web site; sites with few links have a lower probability of being indexed. The most certain and direct approach is to purchase keywords on a search engine; a query with a site's keyword is guaranteed to return the site, normally before those listed by the merit of rank. Once a Web site is discovered by search engines, the methods examined earlier to influence automated search become important, though often the best strategy is to develop and maintain high-quality content to attract and cultivate loyal visitors. Where content quality or time is in shorter supply than money, the paid listing will guarantee that a site is highly ranked by at least one search engine.

Measuring Success

How can a site's owner determine if efforts to attract search engine attention have been a success? Search engines represent the most obvious and direct means to check if and to what extent a specific search engine has indexed a site. Table 1 contains many of the controls needed to limit search to a specific Web site. These same controls used by searchers can also provide feedback to point out search problems with the Web site. Although tests with individual search engines will determine if and how a Web site has been indexed, it will not tell if, why, or how anyone visits. The site server holds the primary information on Web site success in the server access log file. The log holds details about every attempted or successful visit; Table 2 lists the information retained in the Web server access log following the Common Logfile Format. Free and commercial analysis software can produce detailed summaries and graphs of the log; however, the most telling information about search success is contained in the following three fields:

Client Request Line: Contains the page on the server the visitor requested. For visitors arriving from a search engine, this contains the link to the page indexed by the spider.

Server Status Code: Status codes starting with a 2 indicate success; those starting with 4 indicate that the visitor probably encountered a mistake. In Table 2, the "mainpage.htm" page does not exist, earning the visitor a "404 Not Found" response from the Web site.

Referring Site: The visitor reports the site that referred them to this site. In Table 2, the visitor arrived via a reference to "mainpage.htm" made by the "www.food.com" search engine.

Of what value is the access log? Examining the log entries points out errors and successes. Counting the number of visitor requests (i.e., client request line) to each page immediately grades pages on success in attracting visitors and, by their absence, identifies those pages that failed. Investigating the referrer list will show how visitors arrive at a Web site; search engines missing from the list have not indexed the Web site or rank its pages below others. A table of visitor page requests with the referring site will clearly show which search engines successfully found specific pages and can flag pages that create indexing problems for particular search engines. As discussed earlier, some spiders are stopped by frames, only index the first few content lines, or crawl a limited number of links on each site. Pages that are never accessed can indicate indexing problems for the spider or navigation problems for the visitor. Examining the access log file is a good starting point for finding these and other potential search engine and link problems.

Can the log tell us when a complete site, or at least part of a site, is broken? Interpreting the three fields for client request line, server status code, and referring site from the Table 2 example tell us that the "food.com" search engine referred a visitor via an indexed page link to "mainpage.htm" and that link is now broken, as reported by a "404 Not Found" message. A likely reason the link broke is that the Web site was reorganized since the last visit by the "food.com" spider or the page location on the site otherwise changed. Should we inform "food.com" they now have a problem and need to fix their link to point to the new page location? Unreorganizing an active Web site is no solution because other search engines may have already indexed the new site organization. A limited

solution is to invisibly redirect visitors requesting each old page to the new page location through the server configuration or an individual page redirection file for each page moved.

Self-Search

Does the Web site need its own search engine? One can easily provide visitors with site search by placing a link on site pages to a Web search engine. But Web search engines are far from perfect at indexing Web sites, ignoring pages that are important, and that search results focus not on the one site alone but are mixed with results from other sites. Moreover, changes to the site are ignored for the long periods between spider visits.

Web search engines, which bring most new visitors to a Web site, are often poorly suited for searching a site exclusively. However visitors arrive, one authority (Neilsen, 2000) has found that many visitors immediately use search on arrival as the preferred means of locating information and ignore site navigation aides; those visitors need a search engine tailored to the Web site.

What should a designer look for in a search engine? A search engine for a site is comparable in function to a Web search engine but can limit search to the given site. Beyond the raw power required to index a complete Web site, a key capability is to create and search specified branches of the Web site. Indexing every word of the entire site is easy to do but ignores the different reasons visitors search the site and the principle that search works best when the information searched is narrow and fairly homogenous. For example, a university Web site is probably best searched by separating the business and science schools and giving visitors the choice of whether to search the business, science, or all school categories combined. Separating business from science capitalizes upon natural and recognized differences to create a more narrow, homogenous, and recognizable information area to search.

Another issue is the sophistication and flexibility of the search engine. Does it support automatic or manual word stemming, common misspellings, indexing of the HTML tags, synonyms, inclusion and exclusion query operators, and phrase and proximity search?

Two basic options exist for search dedicated to a site: site operated or by retaining an offsite search service. Having someone else handle search is the easier solution but does not necessarily match the owner's needs; handling search oneself can be more flexible but involves more work and expense.

Search Service

Search service companies will index a few pages or an entire Web site, will operate the search engine on their computer, and are in some cases free. The free services generally index a limited number of pages; in return the service places advertising on each search result page. A good service should index pages located on any Web site, index with reasonable promptness whenever the owner chooses, and provide regular summaries of search activity for the site. The main advantage is that someone else maintains the search engine.

The main disadvantages to a search service are the possible continued cost, the possible limit on pages indexed, scheduling the occasional reindexing, the lack of control over the search results or the result page appearance, and that the advertising banners may not impress your site visitors. However, the most serious problem occurs should the service company drastically change policies or technology or go out of business, forcing the Web site search to change.

Reasons for a Private Search Engine

The only compelling reason to operate a private search engine is to benefit site visitors. Fortunately, operating a basic search engine can be relatively easy and many Web server systems include a search engine. Commercially packaged Linux systems come with the ht://Dig engine installed and ready to index the entire server; the owner need only type the "./rundig" command and add a search form link to the site pages.

The main advantages of a private search engine are control of parameters such as indexing depth and access to the information of the logs produced by the search engine during indexing and visitor queries. The following examines common search engine control parameters:

Excerpts: Search results can include text excerpted from a page to help place query keywords in context. Because the excerpting generally takes place during indexing, the searcher needs a little luck for the excerpt to include at least some of his or her query terms. Controlling the size of the excerpt makes it possible to improve the likelihood that the excerpt will contain some query keywords or possibly the entire page.

Indexing: It is important to control the number of words indexed and how often indexing of the site occurs. Although most pages contain fewer than 2000 words or about 7 pages of typed text, indexing spiders with the word limit too low can routinely miss indexing important parts of large pages. Also important to sites with news, pricing, and other frequently updated information is how often indexing occurs and whether it occurs whenever information changes.

Stop Words: Stop words often include numbers and common words such as "the," "computer," "system," and "HTML" and should not be indexed. Words that occur often on many Web site pages should also be included in the stop word list. For example, on a bread Web site, the word "baking" would possess little value in discriminating one page from another.

Measuring Success: Successful searches mean that the visitor follows or "clicks through" to a suggested page. The Web server access log will list the page the visitor follows to, but only the search engine can log the query words entered to find the page; connecting the query words with the referred page measures the effectiveness of search on a Web site. Other measures of search success are whether visitors actually sought the information they were given or immediately searched again, and whether they followed the referred page further to accomplish some task such as buying a

car or registering for a class. Failed queries are also valuable for determining words to add to pages as indicated by the keywords searchers used, expecting to find information, but failing.

Measuring Failure: Failed searches are valuable in identifying what visitors actually want and expect to find on the site. Visitors are sending a clear message if a Web site for programmers sells nothing except JAVA programming tools but 75% of the searches are for FORTRAN. Search engines that log failed searches give insight into possible improvements. For example, queries containing predictable misspellings of FORTRAN can succeed after adding misspellings such as FOTRAN to the engine's synonym list.

The final expert advice on whether a site needs search appears remarkably precise. The site needs search (Powell, 2000) when "The site consists of data such as part numbers, locations, or more than 100 pages" or "The majority of visitors know what they are looking for, how to ask for it, and want to go directly to it".

CONCLUSION

The design of the Web has proven wildly successful in collecting and connecting scattered information but failed to consider finding information on such a large, dispersed network. The success and continued growth of the Web has been fueled by the success of search engines in establishing some degree of coherence and access to the scattered information. Understanding how search engines operate and their limitations can aid a searcher in guiding a search engine to information and a Web site designer in planning a site for search.

This chapter has examined Web information discovery from the three views of the searcher, the search engine, and the Web site. Fundamental reasons, strategies, and techniques employed by searchers for locating high-quality information were presented. Search engine design, strategy, and limitations were examined to provide searchers with some perspective on the comprehensiveness and accuracy of search engine results, along with methods for testing search performance. Search engine and Web site issues affecting the discovery and ranking of information and measures and means for determining search success and disappointment were offered for improving Web site and page design. Reasons for and against local search were also considered.

Search engines represent big businesses that directly profit by attracting and rewarding searchers with information. Web search will continue to evolve and improve as current research matures and is rapidly incorporated into search engine technology.

GLOSSARY

Common log format A standard format for logging and analyzing Web server messages.

Index List of words extracted from pages and the location of each page where the word was extracted. Used for matching query words and locating the pages containing the matches.

Metasearch The fusion of the results from multiple search engines simultaneously searching on the same query.

Page A Web document containing plain text and hypertext markup language for formatting and linking to other pages.

Page rank A system for ranking Web pages where a link from page A to page B increases the rank of page B.

Phrase search A search for documents containing an exact sentence or phrase specified by a user.

Precision The degree to which a search engine matches pages with a query. When all pages are relevant to the query, precision is 100%.

Proximity search A search for pages containing query words within a mutually close proximity.

Query Words given to search engine in order to locate pages containing the same words.

Recall The degree in which a search engine matches relevant pages. When all relevant pages are matched, recall is 100%.

Relevancy The degree to which a page provides the desired information, as measured by the searcher.

Search engine The software that searches an index of page words for query words and returns matches.

Spider The software that locates pages for indexing by following links from one page to another.

Web site A Web location holding and providing access to pages via the World Wide Web.

CROSS REFERENCES

See *Internet Literacy; Internet Navigation (Basics, Services, and Portals); Web Search Technology; Web Site Design.*

REFERENCES

Arasu, A., Cho, J., Garcia-Molina, H., Paepcke, A., & Raghavan, S. (2001). Searching the Web. *ACM Transactions on Internet Technology, 1*(1), 2–43.

Belew, R. (2000). *Finding out about.* New York: Cambridge University Press.

Dreilinger, D., & Howe, A. (1997). Experiences with selecting search engines using metasearch. *ACM Transactions on Information Systems, 15*(3), 195–222.

Hock, R. (2001). *The extreme searcher's guide to Web search engines* (2nd ed.). Medford, NJ: CyberAge Books, Information Today, Inc.

Kleinberg, J. (1999). Authoritative sources in a hyperlinked environment. *Journal of the Association for Computing Machinery, 46*(5), 604–632.

Lawrence, S., & Giles, C. (1999, July 8). Accessibility of information on the Web. *Nature, 400* 107–109.

Lawrence, S., Coetzee, F., Glover, E., Flake, G., Pennock, D., Krovetz, B., Nielsen, F., Kruger, A., & Giles, L. (2000). Persistence of information on the Web: Analyzing citations contained in research articles. In *Proceedings of the Ninth International Conference on Information and Knowledge Management* (pp. 235–242). New York: ACM Press.

Lewis, Mobilio, & Associates (2000). Consumer Daily Question Study, Fall 2000. Retrieved May, 2002, from http://www.keen.com/documents/corpinfo/pressstudy.asp

Mowshowitz, A., & Kawaguchi, A. (2002). Bias on the Web. *Communications of the ACM, 45*(9), 56–60.

Neilsen, J. (2000). Is navigation useful. The alertbox: Current issues in Web usability. Retrieved May, 2002, from http://www.useit.com/alertbox/20000109.html

Page, L., Brin, S., Motwani, R., & Winograd, T. (1998). The PageRank citation ranking: Bringing order to the Web. Technical Report, Stanford, CA: Stanford University.

Powell, T. (2000). *The complete reference: Web design*. Berkeley, CA: Osborne/McGraw-Hill.

Rosenfield, L., & Morville, P. (1998). *Information architecture for the World Wide Web*. Sebastopol, CA: O'Reilly and Associates.

Salton, G., & McGill, M. (1983). *Introduction to modern information retrieval*. New York: McGraw–Hill.

Sonnenreich, W., & Macinta, T. (1998) *Guide to search engines*. New York: Wiley.

Sullivan, D. (2001). *Search engine features for searchers*. Search Engine Watch. Retrieved May, 2002, from http://searchenginewatch.com/facts/ataglance.html

Users don't learn to search better (2001, November 27). *UIEtips*. Retrieved May, 2002, from http://world.std.com/~uieweb/Articles/not_learn_search.htm

Van Boskirk, S., Li, C., Parr, J., & Gerson, M. (2001). Driving customers, not just site traffic. Forrester. Retrieved May, 2002, from http://www.forrester.com/ER/Research/Brief/0,1317,12053,00.html

Web Search Technology

Clement Yu, *University of Illinois at Chicago*
Weiyi Meng, *State University of New York at Binghamton*

INTRODUCTION

The World Wide Web has emerged as the largest information source in recent years. People all over the world use the Web to find needed information on a regular basis. Students use the Web as a library to find references and customers use the Web to purchase various products. It is safe to say that the Web has already become an important part in many people's daily lives.

The precise size of the Web is a moving target as the Web is expanding very quickly. The Web can be divided into the Surface Web and the Deep Web (or Hidden Web) (Bergman, 2000). The former refers to the collection of Web pages that are publicly indexable. Each such page has a logical address called Uniform Resource Locator or URL. It was estimated that the Surface Web contained about 2 billion Web pages in 2000 (Bergman, 2000). The Hidden Web contains Web pages that are not publicly indexable. As an example, a publisher may have accumulated many articles in digital format. If these articles are not placed on the Surface Web (i.e., there are no URLs for them) but they are accessible by Web users through the publisher's search engine, then these articles belong to the Deep Web. Web pages that are dynamically generated using data stored in database systems also belong to the Hidden Web. A recent study estimated the size of the Hidden Web to be about 500 billion pages (Bergman, 2000).

In the past several years, many search engines have been created to help users find desired information on the Web. Search engines are easy-to-use tools for searching the Web. Based on what type of data is searched, there are document-driven search engines and database-driven search engines. The former searches documents (Web pages) while the latter searches data items from a database system through a Web interface. Database-driven search engines are mostly employed for e-commerce applications such as buying cars or books. This chapter concentrates on document-driven search engines only. When a user submits a query, which usually consists of one or more key words that reflect the user's information needs, to a search engine, the search engine returns a list of Web pages (usually their URLs) from the set of Web pages covered by the search engine. Usually, retrieved Web pages are displayed to the user based on how well they are deemed to match with the query, with better-matched ones displayed first. Google (http://www.google.com), AltaVista (http://www.altavista.com), and Lycos (http://www.lycos.com) are some of the most popular document-driven search engines on the Web. The Deep Web is usually accessed through Deep Web search engines (like the publisher's search engine we mentioned earlier). Each Deep Web search engine usually covers a small portion of the Deep Web.

While using a search engine is easy, building a good search engine is not. In order to build a good search engine, the following issues must be addressed. First, how does a search engine find the set of Web pages it wants to cover? After these pages are found, they need to be preprocessed so that their approximate contents can be extracted and stored within the search engine. The approximate representation of Web pages is called the index database of the search engine. So the second issue is how to construct such an index database. Another issue is how to use the index database to determine whether a Web page matches well with a query. These issues will be discussed in the sections Text Retrieval and Search Engine Technology. More specifically, in the former section, we will provide an overview of some basic concepts on text retrieval, including how to build an index database for a given document collection and how to measure the closeness of a document to a query; in the latter, we will provide detailed descriptions about how a search engine finds the Web pages it wants to cover, what are the new ways to measure how well a Web page matches with a query, and what are the techniques to organize the results of a query.

Due to the huge size and the fast expansion of the Web, each search engine can only cover a small portion of the Web. One of the largest search engines on the Web, Google, for example, has a collection of about 3 billion Web pages. However, the entire Web is believed to have more than 500 billion pages. One effective way to increase the search coverage of the Web is to combine the coverage of multiple search engines. Systems that do such combination are called *metasearch engines*. A metasearch

engine can be considered as a system that supports unified access to multiple existing search engines. Combining many Deep Web search engines can be an effective way to enable the search of a large portion of the Deep Web. A metasearch engine does not maintain its own collection of Web pages but it may maintain information about its underlying search engines in order to achieve higher efficiency and effectiveness. In a typical session using a metasearch engine, the user submits a query to the metasearch engine, which passes the query to its underlying search engines; when the metasearch engine receives the Web pages returned from its underlying search engines, it merges these pages into a single ranked list and display them to the user. Techniques that can be used to build efficient and effective metasearch engines are reviewed in the section Metasearch Engine Technology. In particular, we will review techniques for selecting the right search engines to invoke for a given query and techniques for merging results returned from multiple search engines.

This chapter concentrates on the techniques used to build search engines and metasearch engines. Another chapter in this encyclopedia covers Web-searching related issues from the users' and Web page authors' points of view.

TEXT RETRIEVAL

Text (information) retrieval deals with the problem of how to find relevant (useful) documents for any given query from a collection of text documents. Documents are typically preprocessed and represented in a format that facilitates efficient and accurate retrieval. In this section, we provide a brief overview of some basic concepts in classical text retrieval.

The contents of a document may be represented by the words contained in it. Some words such as "a," "of," and "is" do not contain semantic information. These words are called *stop words* and are usually not used for document representation. The remaining words are content words and can be used to represent the document. Variations of the same word may be mapped to the same term. For example, the words "beauty," "beautiful" and "beautify" can be denoted by the term "beaut." This can be achieved by a *stemming program*, which removes suffixes or replaces them by other characters. After removing stop words and stemming, each document can be logically represented by a vector of n terms (Salton and McGill, 1983), where n is the total number of distinct terms in the set of all documents in a document collection.

Suppose the document d is represented by the vector $(d_1, \ldots, d_i, \ldots, d_n)$, where d_i is a number (weight) indicating the importance of the ith term in representing the contents of the document d. Most of the entries in the vector will be zero because most terms do not appear in any given document. When a term is present in a document, the weight assigned to the term is usually based on two factors, namely the *term frequency* (*tf*) factor and the *document frequency* (*df*) factor. The term frequency of a term in a document is the number of times the term appears in the document. Intuitively, the higher the term frequency of a term is, the more important the term is in representing the contents of the document. Consequently, the *term frequency weight* (*tfw*) of a term in a document is usually a monotonically increasing function of its term frequency. The document frequency of a term is the number of documents having the term in the entire document collection. Usually, the higher the document frequency of a term is, the less important the term is in differentiating documents having the term from documents not having it. Thus, the weight of a term with respect to document frequency is usually a monotonically decreasing function of its document frequency and is called the *inverse document frequency weight* (*idfw*). The weight of a term in a document can be the product of its term frequency weight and its inverse document frequency weight, i.e., $tfw * idfw$.

A typical query for text retrieval is a question written in text. It can be transformed into an n-dimensional vector as well, following the same process for transforming documents to vectors described above.

After all documents and a query have been represented as vectors, the query vector can be matched against document vectors so that well-matched documents can be identified and retrieved. Usually, a *similarity function* is used to measure the degree of closeness between two vectors. Let $q = (q_1, \ldots, q_n)$ and $d = (d_1, \ldots, d_n)$ be the vectors of a query and a document, respectively. One simple similarity function is the dot product function, $dot(q, d) = \sum_{i=1}^{n} q_i * d_i$. Essentially, this function is a weighted sum of the terms in common between the two vectors. For a given query, the dot product function tends to favor longer documents over shorter ones. Longer documents have more terms than shorter documents, and as a result, a longer document is more likely to have more terms in common with a given query than a shorter document. One way to remedy this bias is to employ the *Cosine* function, given by $dot(q, d)/(|q| \cdot |d|)$, where $|q|$ and $|d|$ denote, respectively, the lengths of the query vector and the document vector. The Cosine function (Salton and McGill, 1983) between two vectors is really the cosine of the angle between the two vectors. In other words, the Cosine function measures the angular distance between a query vector and a document vector. When the weights in vectors are nonnegative, the Cosine function always returns a value between 0 and 1. It gets the value 0 if there is no term in common between the query and the document (i.e., when the angle is 90°); its value is 1 if the query and the document vectors are identical or one vector is a positive constant multiple of the other (i.e., when the angle is 0°).

Computing the similarity between a query and every document directly is inefficient because many documents do not have any term in common with a given query and computing the similarity of these documents is a waste of effort. To improve the efficiency, an *inverted file index* is created in advance. For each term t_i, an inverted list of the format $[(D_{i1}, w_{i1}), \ldots, (D_{ik}, w_{ik})]$ is generated and stored, where D_{ij} is the identifier of a document containing t_i and w_{ij} is the weight of t_i in D_{ij}, $1 \leq j \leq k$. In addition, a *hash table* is created to locate for each given term the storage location of the inverted file list of the term. These two data structures, namely the inverted file and the hash table, permit efficient calculation of the similarities of all those documents that have positive similarities with any query. Consider a query with m terms. For each query term, the

hash table is used to quickly locate the inverted file list for the term. The m inverted file lists essentially contain all the data needed to calculate the similarity between the query and every document that contains at least one query term.

The retrieval effectiveness of a text retrieval system is often measured by a pair of quantities known as *recall* and *precision*. Suppose, for a given user query, the set of relevant documents in the document collection is known. Recall and precision are defined as

$$\text{recall} = \frac{\text{the number of retrieved relevant documents}}{\text{the number of relevant documents}}$$

$$\text{precision} = \frac{\text{the number of retrieved relevant documents}}{\text{the number of retrieved documents}}.$$

To evaluate the effectiveness of a text retrieval system, a set of test queries is often used. For each query, the set of relevant documents is identified in advance. For each test query, a precision value for each distinct recall value is obtained. Usually only 11 recall values, 0.0, 0.1, . . . , 1.0, are considered. When the precision values at each recall value are averaged over all test queries, an average recall-precision curve is obtained. This curve is used as the measure of the effectiveness of the system.

An ideal text retrieval system should retrieve exactly the set of relevant documents for each query. In other words, a perfect system has recall = 1 and precision = 1 at the same time. In practice, perfect performance is not achievable due to many reasons. For example, a user's needs may be incorrectly or imprecisely specified by the query used and the representation of documents and queries as vectors does not capture their contents completely.

SEARCH ENGINE TECHNOLOGY

A Web search engine is essentially a Web-based text retrieval system for Web pages. However, the Web environment has some special characteristics that make building search engines significantly different from building traditional text retrieval systems. In this section, we review these special characteristics as well as techniques that address/explore these characteristics.

The following are some of the special characteristics of the Web environment.

Web pages are stored at numerous autonomous Web servers. A program known as *Web robot* or *Web spider* is needed to gather them so that they can be processed for later search. Web robots are described below.

Web pages are highly tagged documents; that is, tags that control the presentations of contents on a Web browser and/or define the structures/semantics of the contents are embedded in Web pages. At present, most Web pages are in HTML (hypertext markup language) format, but XML (extensible markup language) documents are making their ways into the Web. Tags in Web pages often convey rich information regarding the terms in these pages. For example, a term appearing in the title of a page or a term highlighted by special font can provide a hint that the term is rather important in reflecting the contents of the page. Traditional text retrieval systems are usually based on plain text documents that are either not or rarely tagged. In Use of Tag Information, we discuss some techniques that utilize these tags to improve retrieval effectiveness.

Web pages are linked to each other. A link from page A to page B allows a Web user to navigate from page A to page B. Such a link also contains several pieces of information that are useful to improve retrieval effectiveness. First, the link indicates a good likelihood that the contents of the two pages are related. Second, the author of page A considers page B to be of some value. Third, the set of words associated with the link, called *anchor terms* of the link, usually provides a short description of page B. In Use of Linkage Information, several methods for utilizing link information among Web pages to improve the search engine retrieval performance will be described.

Another special characteristic of the Web is its size and popularity. A large search engine can index billions of pages and needs to process millions of queries a day. For example, the Google search engine has indexed over 3 billion pages and processes over 27 million queries daily. To accommodate the high computation demand, a large search engine often uses numerous computers with large memory and cache devices, in addition to employing efficient query processing techniques. For example, the inverted file lists for different terms can be stored at different computers. When a multiterm query is received, all computers containing the inverted file list of at least one query term can participate in the evaluation of the query in parallel.

Web Robot

A Web robot (also known as *spider* and *crawler*) is a program for fetching Web pages from remote Web servers. Web robots are widely used to build the Web page collection for a search engine.

The main idea behind the robot program is quite simple. Note that each Web page has a URL that identifies the location of the page on the Web. The program takes one or more seed URLs as input to form an initial URL queue. The program then repeats the following steps until either no new URLs can be found or enough pages have been fetched: (1) take the next URL from the URL queue and fetch the corresponding Web page from its server using the HTTP (hypertext transfer protocol); (2) from each fetched Web page, extract new URLs and add them to the queue.

One of the more challenging problems when implementing a Web robot is to identify all (new) URLs from a Web page. This requires the identification of all possible HTML tags and tag attributes that may hold URLs. While most URLs appear in the anchor tag (e.g., ...), URLs may also appear in other tags. For example, a URL may appear in the option tag as in <option value="URL"...>...</option>, in the area tag (map) as in <area href="URL"...>...</area>, or in the frame tag as in <frame src="URL"...>...</frame>. Frequently a URL appearing in a Web page P does not contain the full path needed to locate the corresponding Web page from a browser. Instead, a partial or relative path is used and a full path must be constructed using the relative path and a base path associated with P. When building a Web robot, it is important to be considerate to

remote servers from which Web pages are fetched. *Rapid fires* (fetching a large number of Web pages from the same server in a short period of time) may overwhelm a server. A well-designed Web robot should control the pace of fetching multiple pages from the same server.

Use of Tag Information

At present, most Web pages are formatted in HTML, which contains a set of tags such as *title* and *header*. Most tags appear in pairs with one indicating the start and the other indicating the end. For example, in HTML, the starting and ending tags for title are <title> and </title>, respectively. Currently, in the context of search engine applications, tag information is primarily used to help computing the weights of index terms.

In Text Retrieval, a method for computing the weight of a term in a document using its term frequency and its document frequency information was introduced. Tag information can also be used to influence the weight of a term. For example, authors of Web pages frequently use emphatic fonts such as boldface, italics, and underscore to emphasize the importance of certain terms in a Web page. Therefore, terms in emphatic fonts should be assigned higher weights. Conversely, terms in smaller fonts should be given lower weights. Other tags such as title and header can also be used to influence term weights. Many well-known search engines such as AltaVista and HotBot have been known to assign higher weights to terms in titles.

A general approach to utilize tags to adjust term weights is as follows (Cutler, Deng, Manicaan, & Meng, 1999). First, the set of HTML tags is partitioned into a number of subsets. For example, the title tag could be a subset by itself, all header tags (i.e., h1, . . . , h6) could form a subset, all list tags (namely "ul" for unordered list, "ol" for ordered list, and "dl" for descriptive list) could be grouped together, and all emphatic tags could form a subset and the rest of the tags can form yet another subset. Next, based on the partition of the tags, term occurrences in a page are partitioned into a number of classes, one class for each subset of tags. For example, all term occurrences appearing in the title form a class. In addition to these classes, two special classes can be formed for each page P. The first contains term occurrences in plain text (i.e., with no tags) and the second contains anchor terms associated with the links that point to the page P. Let n be the number of classes formed. With these classes, for a given page, the term frequency of each term in the page can be represented as a *term frequency vector*: $tfv = (tf_1, \ldots, tf_n)$, where tf_i is the number of times the term appears in the ith class, $i = 1, \ldots, n$. Finally, different importance can be assigned to different classes. Let $civ = (civ_1, \ldots, civ_n)$ be the *class importance vector* such that civ_i is the importance assigned to the ith class, $i = 1, \ldots, n$. With vectors tfv and civ, the traditional term frequency weight formula can be extended into $tf_1 * civ_1 + \cdots + tf_n * civ_n$. This formula takes both the frequencies of the term in different classes and the importance of different classes into consideration. Note that when the civ for the anchor term class is set 0 and all other civ's are set 1, $tf_1 * civ_1 + \cdots + tf_n * civ_n$ becomes tf, the total frequency of the term in a page.

An interesting issue is how to find the optimal class importance vector that can yield the highest retrieval effectiveness. One method is to find an optimal or near optimal civ experimentally based on a test bed (Cutler et al., 1999). The test bed contains a Web page collection and a set of queries; for each query, the set of relevant pages is identified. Then different civs can be tested based on the test bed and the civ that yields the highest retrieval effectiveness can be considered as an optimal civ. In practice, the number of different civs may be too large to permit all civs to be tested. Instead, heuristic algorithms such as a genetic algorithm may be employed to find an optimal or near optimal civ efficiently.

Use of Linkage Information

There are several ways to make use of the linkage information between Web pages to improve the retrieval quality of a search engine. One method is to use all anchor terms associated with links that point to page B to represent the contents of B. Since these terms are chosen by human authors for the purpose of describing the contents of B, they are of high-quality terms for indicating the contents of B. Many search engines (e.g., Google) use anchor terms to index linked pages (e.g., page B). The most well-known uses of the linkage information in search engines treat links as votes to determine the importance of Web pages. One method (the PageRank method) tries to find the overall importance of each Web page regardless of the contents of the page, and another method (the Hub-and-Authority method) tries to identify pages that provide good content pages (authoritative pages) with respect to a given topic as well as good reference pages (hub pages) to these good content pages. These two methods are described below.

PageRank Method

The Web can be viewed as a gigantic directed graph $G(V, E)$, where V is the set of pages (vertices) and E is the set of links (directed edges). Each page may have a number of outgoing edges (forward links) and a number of incoming edges (backlinks). As mentioned at the beginning of Search Engine Technology, when an author places a link in page A to point to page B, the author shows that he/she considers page B to be of some value. In other words, such a link can be viewed as a vote for page B. Each page may be pointed to by many other pages and the corresponding links can be aggregated in some way to reflect the overall importance of the page. For a given page, *PageRank* is a measure of the relative importance of the page on the Web, and this measure is computed based on the linkage information (Page, Brin, Motwani, & Winograd, 1998). The following are the three main ideas behind the definition and computation of the PageRanks of Web pages.

Pages that have more backlinks are likely to be more important. Intuitively, when a page has more backlinks, the page is considered to be of some value by more Web page authors. In other words, the importance of a page should be reflected by the popularity of the page among all Web page authors. For example, the home page of Yahoo! (http://www.yahoo.com) is probably one of the

most linked pages on the Web and it is certainly one of the most important pages on the Web.

The importance of a page increases if it is pointed to by more important pages. Suppose page A and page B are pointed to by two sets of pages S_a and S_b, respectively, and the two sets have the same cardinality. If each page in S_a is more important than the corresponding page in S_b and they have the same number of outgoing pages (see the next paragraph for the reason why the same number of outgoing pointers is needed), then page A is more important than page B. Intuitively, important pages are likely to be published by important authors or organizations and the endorsement of these authors/organizations should have more weight in determining the importance of a page. As an example, if a page is pointed to by the home page of Yahoo!, then the page is likely to be important. To summarize, the importance of a page should be a weighted popularity measure of the page.

When a page has more forward links, the page has less influence over the importance of each of the linked pages. The previous description indicates that the importance of a page may be propagated to its child pages. If a page has multiple child pages, then these pages should share the importance of the parent page. In other words, if a page has more child pages, then it can only propagate a smaller fraction of its importance to each child page.

From the above discussion, it can be seen that the PageRanks of Web pages need to be computed in a recursive manner. The PageRank of a page u is defined more formally as follows. Let F_u denote the set of pages that are pointed to by u and B_u denote the set of pages that point to u. For a set X, let $|X|$ denote the number of items in X. The PageRank of u, denoted $R(u)$, is defined by

$$R(u) = \sum_{v \in B_u} \frac{R(v)}{|F_v|}. \quad (1)$$

Note how the above equation incorporates the three ideas we discussed earlier. First, the sum reflects the idea that more backlinks can lead to a larger PageRank. Second, having $R(v)$ in the numerator indicates that the PageRank of u is increased more if page v is more important (i.e., has larger $R(v)$). Third, using $|F_v|$ in the denominator implies that the importance of a page is evenly divided and propagated to each of its child pages. Also note that Equation (1) is recursive. The PageRanks of Web pages can be computed as follows. First, an initial PageRank is assigned to all pages. Let N be the number of Web pages. Then $1/N$ could be used as the initial PageRank of each page. Next, Equation (1) is applied to compute the PageRanks in a number of iterations. In each iteration, the PageRank of each page is computed using the PageRanks of its parent pages in the previous iteration. This process is repeated until the PageRanks of the pages converge within a given threshold.

The PageRanks of Web pages can be used to help retrieve better Web pages in a search engine environment. In particular, the PageRanks of Web pages can be combined with other, say content-based, measures to indicate the overall relevance of Web pages with respect to a given query. For example, suppose for a given query, the similarity of a page p with the query is $sim(p)$. From this similarity and the PageRank of p, $R(p)$, a new score of the page with respect to the query can be computed by $w * sim(p) + (1 - w) * R(p)$, where 0 <w< 1. This formula emphasizes not only how similar the page is to the query but also how important the page is. As a result, a search engine that employs such a formula to rank Web pages tends to favor more important pages that have similar similarities. Utilizing PageRank in retrieval has an anti-spamming effect. Some Web page authors insert many popular but irrelevant words in a page to boost its chance of being retrieved by search engines. When PageRank is used, the search engine no longer relies on only words appearing in a page to determine its potential relevance.

One of the most distinctive features of the Google search engine from other Web search engines is its incorporation of the PageRanks of Web pages, together with other factors such as similarities and proximity information, into the Web page selection process when processing a user query.

Hub-and-Authority Method

PageRank measures the global relative importance of Web pages but the importance is topic independent. Although the global importance is very useful, we often also need topic-specific importance. For example, a movie lover is more likely to be interested in important movie pages. A topic-specific important page can be considered to be an authoritative page with respect to a given topic. The question here is how to measure and find authoritative pages for a given topic. Kleinberg (1998) proposed using authority scores to measure the importance of a Web page with respect to a given topic. He also suggested a procedure for finding topic-specific important pages. The main ideas of this approach are reviewed here.

According to Kleinberg (1998), Web pages may be conceptually classified into two types: *authoritative pages* and *hub pages*. The former contains informative contents on some topic and the latter has links to authoritative pages. It is observed that for a given topic, a good authoritative page is often pointed to by many good hub pages and a good hub page often has links to good authoritative pages. Good authoritative pages and good hub pages related to the same topic often reinforce each other. It is possible for the same page to be both a good authoritative page and a good hub page on the same topic.

In the search engine context, a topic can be defined by a user query submitted to a search engine. The procedure proposed in Kleinberg (1998) for finding good authoritative and hub pages can be summarized into the following three steps. Let q be a user-submitted query.

Process the query by submitting it to a regular (similarity-based) search engine. Let S denote the set of documents returned by the search engine. The set S is called the *root set*. In order to find good authoritative pages with respect to a query, the size of S should be reasonably large, say in the hundreds.

Expand the root set S into a larger set T called the *base set*. T is expanded from S by including any page that points to a page in S or is pointed to by a page in S. In other words, T contains all pages in S as well as all the parent pages and child pages of the pages in S. In order to compute T

from S quickly, the search engine can save the link structure of the Web in advance. While a Web page typically has a small number of child pages, it may have an extremely large number of parent pages. For example, the home page of Yahoo! may be linked by millions of pages. In order to limit the size of the base set, a threshold, say 20, can be used such that at most 20 parent pages of each page in S are included in T. Similar restriction may also be applied to the inclusion of child pages. Some heuristics may be applied to choose parent pages for inclusion. One heuristic is to select those parent pages whose anchor terms associated with the links contain terms in the query. The base set for a given query typically contains thousands of pages related to the query. The last step is to identify the most authoritative pages with respect to the query from the base set.

Compute the authority score and hub score of each page in the base set T. For a given page p, let $a(p)$ and $h(p)$ be the authority and hub scores of p, respectively. Initially, $a(p) = h(p) = 1$ for each page p. For pages p and q, let (p, q) denote a link from p to q. The computation is carried out in a number of iterations. In each iteration, two basic operations and two normalizations are executed for each page. The two operations are called Operation I and Operation O.

Operation I: Update each $a(p)$ to be the sum of the current hub scores of Web pages in the base set that point to p. More precisely,

$$a(p) = \sum_{q:(q,p)\in E} h(q),$$

where E is the set of links between pages in the base set T.

Operation O: Update each $h(p)$ to be the sum of the current authority scores of Web pages in the base set linked from p. More precisely,

$$h(p) = \sum_{q:(p,q)\in E} a(q).$$

After all authority and hub scores have been updated in the current iteration, normalize each authority score and each hub score as

$$a(p) = \frac{a(p)}{\sqrt{\sum_{q\in T}[a(q)]^2}}, \quad h(p) = \frac{h(p)}{\sqrt{\sum_{q\in T}[h(q)]^2}}.$$

The above computation process is repeated until the scores converge.

After all scores are computed, the Web pages in the base set are sorted in descending authority score and the pages with top authority scores are displayed to the user.

The above method for calculating authority and hub scores treats all links the same. In reality, some links may be more useful than other links in identifying authoritative pages with respect to a given topic. Two cases are considered below.

Two types of links can be distinguished based on whether the two pages related to a link are from the same domain, where the domain name of a page is the first-level string of its URL (i.e., the part before the first "/"). A link between pages with different domain names is called a *transverse link* and a link between pages with the same domain name is called an *intrinsic link*. It can be argued that transverse links are more significant than intrinsic links for two reasons. First, intrinsic links are sometimes used for presentation purposes, i.e., breaking a large document into smaller linked pieces to facilitate browsing. Second, intrinsic links can be considered to be self-referencing whose significance should be lower than references made by others. One way to handle intrinsic links is to simply discard them (Kleinberg, 1998). Another method is to give a lower weight to intrinsic links (Chakrabarti et al., 1999).

Recall that a link often has a set of associated anchor terms. It can be argued that if the anchor terms of a link contain terms in the query (topic), then the likelihood that the page pointed to by the link is an authoritative page with respect to the topic is increased. In general, a vicinity of a link can be defined to include terms within a certain distance (say 50 characters) on both sides on the link. Then the weight associated with a link can be defined as an increasing function of the number of terms in the vicinity of the link that appear in the query and this weight can be incorporated into the computation of authority and hub scores (Chakrabarti et al., 1998).

Use of User Profiles

Successful information finding on the Web is sensitive to the background interests of individual users. For example, for the query "apple," a person with a history of retrieving documents in the computer area is more likely to be interested in information related to "Apple computer" while a person interested in food probably would like to see pages that consider apple as a fruit. Personalized search is an important method for improving the retrieval effectiveness of a search engine. The background information of each user (i.e., user profile) can be collected in a number of ways such as through the use of bookmarks and cookies.

Parts of user profiles can also be generated from implicit user feedback. When a user submits a query to a search engine, the user may have some of the following behaviors or reactions regarding the returned Web pages:

Click some pages in certain order while ignoring others.
Read some clicked pages for a longer time than some other clicked pages.
Save/print certain pages.

If a user saves/prints a page or spends a significant amount time on a page (before clicking another page), then the page can be considered as useful to the user. Information, such as significant terms, extracted from the queries submitted by a user and from the pages considered to be useful to the user can be used, together with possibly other information about the user such as bookmarks and cookies, to form a profile for the user. In general, a user may have multiple profiles corresponding to the different interests of the user and these profiles may evolve with time.

User profiles may be utilized in a number of ways to improve the retrieval effectiveness of a search engine as discussed below.

User profiles can be used to help determine the meaning of a query term. If a user submits a query with a single term "bank" and the user has a profile on environment but no profile on finance, then it is likely that the current usage of this term is like in "river bank" rather than in "investment bank." Web users often submit short queries to search engines. A typical query has about 2.3 terms and about 30% of all queries have just one term. For these short queries, correctly determining the meanings of query terms can help retrieve relevant Web pages.

User profiles can be utilized to perform *query expansion*. When an appropriate profile can be identified to be closely related to a query, then terms in the profile may be added to the query such that a longer query can be processed. In text retrieval, it is known that longer queries tend to return better-matched documents because they are often more precise in describing users' information needs than short queries.

User profiles can be used to filter initial results. When a user query is received by a search engine, a list of results based on only the query but not any profile of the user can be obtained first. These results can then be compared with the profiles of the user to help identify Web pages that are more likely to be useful to this particular user.

User profiles of one user can be used to help find useful pages for another user. Part of a user profile may include what queries have been submitted by a user and what pages have been considered as useful for each query. When a user u_1 submits a new query q, it is possible to find another user u_2 such that the profiles of the two users are similar and user u_2 has submitted query q before. In this case, the search engine may rank highly the pages that were identified to be useful by u_2 for query q in the result for u_1. Furthermore, from the profiles of all users, it is possible to know how many users have considered a particular page to be useful (regardless of for what queries). Such information can be used to create a *recommender system* in the search engine environment. Essentially, if a page has been considered to be useful by many users for queries similar to a newly received query, then the page is likely to be useful to the new query and should be ranked high in the result. The DirectHit search engine (http://www.directhit.com) has incorporated the principles of recommender systems.

Result Organization

Most search engines organize retrieved results (including URLs and some short descriptions known as *snippets*) in descending order of their estimated desirabilities with respect to a given query. The *desirability* of a page to a query could be approximated in many different ways such as the similarity of the page with the query, a combined measure including similarity and rank of the page, or the authority score of the page. Some search engines, such as FirstGov (http://www.firstgov.gov) and Northern Light (http://www.northernlight.com) also provide the estimated desirabilities of returned pages while some, such as AltaVista and Google, do not provide such information.

Some search engines organize their results into groups such that pages that have certain common features are placed into the same group. Such an organization of the results, when meaningful labels (annotations) are assigned to each group, can facilitate users for identifying useful pages from the returned results. This is especially useful when the number of pages returned for a query is large. The Vivisimo search engine (http://www.vivisimo.com) and the DynaCat system (Pratt, Hearst, & Fagan, 1999) organize retuned results for each query into a hierarchy of groups. In general, the issues that need to be addressed when implementing an online result-clustering algorithm include: (1) What information (titles, URLs, snippets versus the full documents) should be used to perform the clustering? While more information may improve the quality of the clusters, using too much information may cause long delays for users. (2) How to cluster? A large number of text clustering algorithms exist. (3) How to come up with labels that are meaningful descriptions of each group? (4) How to organize the groups? They could be linearly ordered or hierarchically ordered. In the former case, what should be the linear order? In the latter case, how to generate the hierarchy? (5) How to order pages in each cluster? Many of the issues are still being actively researched.

Study indicates that clustering/categorizing search results is effective in helping user identify relevant results (Hearst & Pedersen, 1996), especially when user queries are short. Short queries often result in diverse results because short queries can have different interpretations. When results are organized into multiple clusters, results corresponding to the same interpretation tend to fall in the same cluster. In this case, when clusters are appropriately annotated, finding relevant results becomes much easier.

METASEARCH ENGINE TECHNOLOGY

A metasearch engine provides a way to access multiple existing search engines with ease. One of the most significant benefits of metasearch engines is its ability to combine the coverage of many search engines. In particular, by employing many search engines for the Deep Web, a metasearch engine can be an effective tool for quickly reaching a large portion of the Deep Web. In this section, we provide an overview of the metasearch engine technology. First, a reference software component architecture of a metasearch engine is introduced. Then in Database Selection, techniques that identify what search engines are likely to contain useful results for a given query are discussed. The set of Web pages that can be searched by a search engine is the Web page database of the search engine. Therefore, the search engine selection problem is also known as the database selection problem. In Collection Fusion, methods that determine what pages from selected search engines should be retrieved and how the results from different search engines should be merged are reviewed.

Software Component Architecture

A reference software component architecture of a metasearch engine (Meng, Yu, & Liu, 2002) is illustrated in Figure 1. The numbers associated with the arrows indicate the sequence of steps for processing a query. More

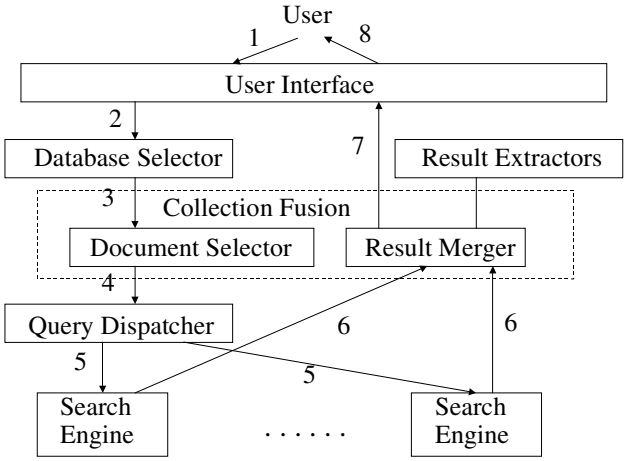


Figure 1: Metasearch software component architecture.

details regarding each software component are described below.

Database Selector. If the number of local search engines in a metasearch engine is small, then it would be reasonable to send each user query submitted to the metasearch engine to all the local search engines. However, if the number is large, say in the hundreds, then sending each query to all local search engines will be an inefficient strategy because most local search engines will be useless with respect to any given query. As an example, suppose only the 10 best-matched pages are needed for a query. Clearly, the 10 desired pages are contained in no more than 10 search engines. This means that if there are 500 local search engines, then 490 of them are useless with respect to this query. Sending a query to useless search engines may cause serious inefficiencies. For example, transmitting the query to useless search engines from the metasearch engine and transmitting useless retrieved pages from these search engines to the metasearch engine would cause wasteful network traffic. As another example, when a query is evaluated at useless search engines, system resources at these local systems would be wasted. Therefore, it is important to send each user query to only potentially useful search engines for processing. The problem of identifying potentially useful search engines to invoke for a given query is known as the *database selection problem*. The software component *database selector* is responsible for performing database selection. Selected database selection techniques will be discussed in Database Selection.

Collection Fusion. When searching from multiple document databases, *collection fusion* is a method for providing transparency to multiple databases. In the metasearch engine context, a collection fusion method determines what Web pages should be retrieved from each selected search engine and how the retrieved Web pages from multiple search engines should be merged into a single result list. In other words, collection fusion consists of a document selection module (*document selector*) and a result merge module (*result merger*). More details about these two modules are provided below.

Document Selector. For each search engine selected by the database selector, *document selector* determines what pages to retrieve from the document database of the search engine. The objective is to retrieve, from each selected local search engine, as many potentially useful pages as possible, and as few useless pages as possible. When more useless pages are returned from a search engine, greater effort would be needed by the metasearch engine to identify potentially useful ones.

Result Merger. After the results from selected local search engines are returned to the metasearch engine, the *result merger* combines the results into a single ranked list. The top m pages in the list are then returned to the user through the user interface, where m is the number of pages desired by the user. A good result merger should rank all returned pages in descending order of their desirabilities.

Result Extractor. One technical issue related to result merging is result extraction. When search results are returned by a search engine, they are grouped into one or more result pages, which contain the URLs and possibly some snippets of retrieved Web pages. Each result page is a dynamically generated HTML file. Usually, in addition to the URLs of retrieved pages, a result page also contains URLs unrelated to the user query. These unrelated URLs include URLs for advertisement pages and service pages. Therefore, the URLs of retrieved pages need to be correctly extracted from the HTML file of each result page. Since different search engines use different ways to organize their result, a separate *result extractor* needs to be created for each local search engine. Result extraction will not be discussed further in this chapter.

Different methods for performing document selection and result merging will be discussed in Collection Fusion.

Query Dispatcher. After a local search engine has been selected to participate in the processing of a user query, the *query dispatcher* establishes a connection with the server of the search engine and passes the query to it. HTTP is used for the connection and data transfer (sending queries and receiving results). In general, different search engines have different requirements on the HTTP request method, such as the GET method or the POST method, and the query format, such as the specific query box name. Therefore, the query dispatcher consists of many connection programs (wrappers), one for each local search engine. In addition, the query sent to a particular search engine may or may not be the same as that received by the metasearch engine. In other words, the original user query may be modified to a new query before being sent to a search engine. For vector space queries, query modification is usually not needed. Two possible types of modifications are as follows. First, the relative weights of query terms in the original user query may be changed before the query is sent to a local search engine. The change could be accomplished by repeating some query terms an appropriate number of times as the weight of a term is usually an increasing function of its frequency. Such a modification on query term weights could be useful to influence the ranking of the retrieved Web pages from the

local search engine in a way as desired by the metasearch engine (Meng et al., 2002). Second, the number of pages to be retrieved from a local search engine may be different from that desired by the user. For example, suppose as part of a query, a user of the metasearch engine indicates that m Web pages are desired. The document selector may decide that k pages should be retrieved from a particular local search engine. In this case, the number k, usually different from m, should be part of the modified query to be sent to the local search engine. Note that not all search engines on the Web support the specification of the desired number of pages by a user. For these search engines, the second type of query modification is not possible. Query dispatch will not be discussed further in this chapter.

Database Selection

As we explained in Software Component Architecture, when a metasearch engine receives a query from a user, the database selector is invoked to select local search engines likely to contain useful Web pages for the query. To enable database selection, some characteristic information representing the contents of the document database of each search engine needs to be collected and made available to the database selector. The characteristic information about a database will be called the *representative* of the database in this chapter. Many database selection techniques have been proposed, and these techniques can be classified into the following three categories (Meng et al., 2002):

Rough representative approaches. In these approaches, the representative of a database contains only a few selected key words or paragraphs. Clearly, rough representatives can only provide a very general description about the contents of databases. Consequently, database selection techniques based on rough representatives are not very accurate in estimating the true usefulness of each database with respect to a given query. Rough representatives are often manually generated.

Statistical representative approaches. Database representatives using these approaches have detailed statistical information about the document databases. Typically, statistics for each term in a database such as the *document frequency* of the term and the average weight of the term in all documents having the term are collected. While detailed statistics allow more accurate estimation of database usefulness with respect to any user query, more effort is needed to collect them and more storage space is needed to store them.

Learning-based approaches. As the databases of different local search engines are different, they are not equally useful for a given query. Learning-based approaches learn the knowledge regarding which databases are likely to return useful pages to what types of queries from past retrieval experiences. Such knowledge is then used to determine the usefulness of databases for each new query. For these approaches, the representative of a database is simply the knowledge indicating the past performance of the database with respect to different queries.

The main appeal of rough representative approaches is that the representatives can be obtained relatively easily, and they require little storage space. If all local search engines in a metasearch engine are highly specialized with diversified topics such that the contents of their databases can be easily summarized and differentiated, then these approaches may work reasonably well. On the other hand, it is unlikely that the short representative of a database can summarize the contents of the database sufficiently comprehensively, especially when the database contains Web pages of diverse topics. Therefore, these approaches can easily miss potentially useful databases for a query when performing database selection. A widely used method to alleviate this problem is to involve users in the database selection process. For example, in ALIWEB (Koster, 1994) and WAIS (Kahle & Medlar, 1991), users will make the final decision on which databases to use based on the preliminary selections made by the database selector. In another system that employs rough database representatives, Search Broker (Manber & Bigot, 1997), user queries are required to contain the subject areas for the queries. Users often do not know all local search engines well. As a result, their contribution in database selection is limited. In general, rough representative approaches are inadequate for large-scale metasearch engines. Rough representative approaches will not be discussed further in this chapter. For the rest of this section, we concentrate on the other two types of approaches.

Statistical Representative Approaches

A statistical representative of a database typically takes every distinct term in every page in the database into consideration. Usually, one or more pieces of statistical information for each term are kept in the representative. As a result, database selection techniques using statistical representatives are more likely to be able to detect the existence of individual potentially useful pages in a database for any given query. A large number of approaches based on statistical representatives have been proposed (Meng et al., 2002). In this chapter, we describe two of these approaches.

CORI Net Approach. CORI Net (Collection Retrieval Inference Network (Callan, Lu, & Croft, 1995)) is a research system for retrieving documents from multiple document collections. Each collection corresponds to a database in a metasearch engine environment. Let $t_1, \ldots, t_n$ be all the distinct terms in all collections in the system. The representative of each collection C conceptually consists of a set of triplets (t_i, df_i, cf_i), $i = 1, \ldots, n$, where cf_i is the *collection frequency* of term t_i (i.e., the number of collections that contain t_i) and df_i is the *document frequency* of t_i in C. If a particular term, say t_j, does not appear in C, then $df_j = 0$ for C and the triplet (t_j, df_j, cf_j) needs not to be kept. Note that the collection frequency of a term is a system-wide statistics and only one cf needs to be kept for each term in the system for collection selection.

For a given query q, CORI Net ranks collections using a technique originally proposed to rank documents in

the INQUERY document retrieval system (Callan, Croft, & Harding, 1992). This technique is known as the *inference network* approach. When ranking collections, CORI Net conceptually treats each collection representative as a (super)document. Thus, the set of all representatives forms a collection of superdocuments. Let D denote this collection of superdocuments. Consider collection C of regular documents and a term t in C. The df_i of t can be treated as the *term frequency* of t in the superdocument of C. Similarly, the cf_i of t can be treated as the *document frequency* of t in D. This means that we have the term frequency and document frequency of each term in each superdocument. With the representative of each collection being treated like a superdocument, the problem of ranking collections has been reduced to the problem of ranking (super)documents. At this point, any term frequency- and document frequency-based term-weighting scheme may be utilized to compute the weight of each term in each superdocument. As a result, all superdocuments can be represented as vectors of terms with weights. Furthermore, any vector-based similarity function such as those discussed under Text Retrieval may be used to compute the similarities between a given query and any superdocument, and these similarities can be used as the ranking scores of collections for the query.

CORI Net employs an inference network-based probabilistic approach to rank superdocuments for a given query. In this approach, the ranking score of a collection with respect to query q is an estimated belief that the collection contains useful documents to the query. The belief is the combined probability that the collection contains useful documents due to each query term. Suppose query q contains k terms $t_1, \ldots, t_k$. Let N be the number of collections under consideration. Let df_{ij} be the document frequency of the jth term in the ith collection C_i and cf_j be the collection frequency of the jth term. The belief that C_i contains useful documents due to the jth query term is estimated by

$$p(t_j \mid C_i) = a_1 + (1 - a_1) \cdot T_{ij} \cdot I_j,$$

where

$$T_{ij} = a_2 + (1 - a_2) \cdot \frac{df_{ij}}{df_{ij} + K},$$

$$I_j = \log\left(\frac{N + 0.5}{cf_j}\right) \Big/ \log(N + 1.0).$$

In the above, T_{ij} is a formula employed by CORI Net for computing the term frequency weight of the jth term in the superdocument of C_i, I_j is for computing the inverse document frequency weight of the jth term, a_1 and a_2 are constants and K is a parameter related to the size of collection C_i. Appropriate values for a_1, a_2, and K can be determined empirically by experiments (Callan et al., 1995). Let $p(q \mid t_j)$ denote the belief that term t_j represents query q. One way to estimate $p(q \mid t_j)$ is to use the weight of t_j in q. Finally, the belief that collection C_i contains documents useful to query q can be estimated to be

$$r_i = p(q \mid C_i) = \sum_{j=1}^{k} p(q \mid t_j) \cdot p(t_j \mid C_i)$$

A nice feature of the CORI Net approach is that the same method can be used to rank documents for a query as well as to rank collections for a query. More information about the CORI Net approach can be found in Callan (2000).

Most Similar Document Approach. One useful measure for ranking databases is the global similarity of the most similar document in a database for a given query. The global similarity may incorporate other usefulness measures of a page such as the PageRank in addition to the usual similarity between a query and the page. Theoretically, documents should be retrieved in descending order of some global measure to achieve optimal retrieval, if such a measure truly indicates the degree of relevance. The global similarities of the most similar documents in all databases can be used to rank databases optimally for a given query for retrieving the m most similar documents across all databases. Suppose there are M local databases $D_1, D_2, \ldots, D_M$ and the m most similar documents from these databases are desired. Intuitively, it is desirable to order these databases in such a way that the first k databases collectively contain the m most similar documents and each of these k databases contains at least one of these documents for some integer k. The following definition introduces the concept of an optimal database order for a given query (Yu, Meng, Liu, Wu, & Rishe, 1999).

Definition. For a given query q, databases $D_1, D_2, \ldots, D_M$ are said to be optimally ranked in the order $[D_1, D_2, \ldots, D_M]$ if there exists an integer k such that $D_1, D_2, \ldots, D_k$ contain all of the m most similar documents and each D_i, $1 \le i \le k$, contains at least one of the m most similar documents.

Clearly, if databases are optimally ranked for a query, then it is sufficient to search the first k databases only. Let $msim(q, D)$ be the global similarity of the most similar document in database D with the query q. It can be shown (Yu et al., 1999) that if databases are ranked in descending order of their $msim(q, D)$'s, then these databases are optimally ranked with respect to q. This theory cannot be applied as is because it is not practical to search all databases, find the global similarities of the most similar documents in these databases, and then rank them for each query. The solution to this problem is to estimate $msim(q, D)$ for any database D using the representative of D stored in the metasearch engine.

One method for estimating $msim(q, D)$ is as follows (Yu et al., 1999). Two types of representatives are used. In the global representative for all local databases, the global inverse document frequency weight ($gidf_i$) is kept for each distinct term t_i. There is a local representative for each local database D. For each distinct term t_i in D, two quantities, mnw_i and anw_i, are kept, where mnw_i and anw_i are the *maximum normalized weight* and the *average normalized weight* of term t_i, respectively. If d_i is the weight of t_i in a document d, then $d/|d|$ is the normalized weight of t_i in d, where $|d|$ is the length of d. For term t_i in database D, mnw_i and anw_i are defined to be the maximum and the average of the normalized weights of t_i in all documents in D, respectively. Let $(q_1, \ldots, q_k)$ be the vector of query q.

Then $msim(q, D)$ can be estimated by the formula

$$msim(q, D) = \max_{1 \le i \le k} \left\{ q_i * gidf_i * mnw_i + \sum_{j=1, j \neq i}^{k} q_j * gidf_j * anw_j \right\} \Big/ |q|. \quad (2)$$

Intuitively, if document d is the most similar document in a database D, then one of the terms in d, say t_i, is likely to have the maximum normalized weight of t_i in all documents in D, while each of the other query terms is expected to take the average normalized weight of the corresponding term. Since any one of the k query terms in d could have the maximum normalized weight, the maximum is taken over all terms t_i. Finally, normalization by the query length, $|q|$, yields a value less than or equal to 1. As $|q|$ is common to all $msim(q, D)$'s, it can be dropped from the above formula without affecting the ranking of databases.

It was observed in (Wu, Meng, Yu, & Li, 2001) that for a given term, its maximum normalized weight is typically two or more orders of magnitude larger than its average normalized weight. This is mainly because the average normalized weight of a term is computed over all documents, including those that do not contain the term. Meanwhile, it can be observed that for most user queries, all query terms have the same *tf* weight, that is, $q_i = q_j$, $i \neq j$. This is because in a typical query, each term appears the same number of times, namely once. From these two observations, we can see that in Equation (2), for a typical query, $gidf_i * mnw_i$ is likely to dominate

$$\sum_{j=1, j \neq i}^{k} gidf_j * anw_j,$$

especially when the number of terms, k, in the query is small (according to Kirsch, 1998, the number of terms in a search engine query is 2.3 on the average). This means that the rank of database D with respect to a given query q is largely determined by $\max_{1 \le i \le k} \{q_i * gidf_i * mnw_i\}$. This leads to the following more scalable formula for ranking databases (Wu et al., 2001): $\max_{1 \le i \le k} \{q_i * am_i\}$, where $am_i = gidf_i * mnw_i$ is the *adjusted maximum normalized weight* of term t_i in D. This formula requires just one quantity, namely am_i, to be kept in the database representative for each distinct term in the database.

Constructing Database Representatives. Statistical database selection methods depend on the availability of detailed statistical information about the terms in the document collection of each search engine. Cooperative search engines may provide desired statistical information about their document collection to the metasearch engine. For uncooperative search engines that follow a certain standard, say the proposed STARTS standard (Gravano, Chang, Garcia-Molina, & Paepcke, 1997), the needed statistics may be obtained from the information that can be provided by these search engines such as the document frequency and the average document term weight of any query term. For uncooperative search engines that do not follow any standard, their representatives may have to be extracted from sampled documents (e.g., Callan, 2000; Callan, Connell, & Du, 1999).

Learning-Based Approaches

The usefulness of a database for different kinds of queries can be learned by examining the returned documents for past queries. The learned knowledge can then be used to select potentially useful databases to search for new queries. This is the idea behind all learning-based database selection techniques. The learning may be conducted in a number of ways. For example, carefully chosen training queries can be used. Ideally, training queries should reflect real user queries and cover the diversified contents of all local databases. Learning based on training queries is usually conducted offline. The knowledge obtained from training queries may become outdated due to the changes of database contents and query patterns. Another way to carry out learning is to use real user queries. In this case, the metasearch engine keeps track of user queries, and for each query, the documents implied by the user to be useful (e.g., the user spends much time with a page), and finally, from which database each implied useful document is retrieved. When real user queries are used for learning, the knowledge base can be updated continuously. Clearly, it may take a while for the metasearch engine to accumulate sufficient knowledge to enable meaningful database selection. It is possible for a metasearch engine to utilize both training queries and real user queries for learning. In this case, training queries are used to obtain the initial knowledge base while real user queries are used to update the knowledge base. Two learning-based database selection methods are reviewed below. The first method uses only real user queries while the second method uses both training queries and real user queries.

SavvySearch Approach. SavvySearch (http://www.search.com, acquired by CNET in 1999) conducts learning based on real user queries and the reactions of real users to retrieved pages. In SavvySearch (Dreilinger & Howe, 1997) the ranking score of a local search engine for a given query is computed based on the past retrieval performance related to queries that contain terms in the new query. More specifically, a weight vector $(w_1, \ldots, w_m)$ is maintained for each local search engine by the metasearch engine, where w_i corresponds to the ith term in the database of the search engine. All weights are zero initially. Consider a user query with k terms. Suppose t_i is one of the query terms and for this query, database D is selected to retrieve documents. Then the weight w_i of t_i for D is adjusted according to the retrieval result. If no document is returned from D, then w_i is reduced by $1/k$, indicating that each of the k query terms contributed equally to the poor result. On the other hand, if at least one returned document is clicked by the user, showing the interest of the user to the document, then w_i is increased by $1/k$, indicating that each of the k query terms is equally responsible for the good result. In other cases, w_i is left unchanged. Over time, if a database has a large positive weight for term t_i, then the database is considered to have responded well to term t_i in the past. Conversely, if a large negative weight is recorded for t_i for a database, then the database has responded poorly to t_i in the past.

An interesting feature of SavvySearch is that its database selection algorithm combines both content-based selection and performance-based selection. Most database selection methods employ content-based selection only. In contrast, performance-based selection method takes into consideration information such as the speed and the connectability of each local search engine when performing database selection. SavvySearch keeps track of two types of performance-related information for each search engine (Dreilinger & Howe, 1997). The first is h, the average number of documents returned for the most recent five queries sent to the search engine, and the other is r, the average response time for the most recent five queries. If h is below a threshold T_h (the default value is 1), then a penalty $p_h = (T_h - h)^2/T_h^2$ for the search engine is applied. Similarly, if the average response time r is greater than a threshold T_r (the default is 15 seconds), then a penalty $p_r = (r - T_r)^2/(r_o - T_r)^2$ is computed, where $r_o = 45$ (seconds) is the maximum allowed response time before a timeout.

For a new query q with terms $t_1, \ldots, t_k$, SavvySearch uses the following formula to compute the ranking score of database D,

$$r(q, D) = \frac{\sum_{i=1}^{k} w_i \cdot \log(N/cf_i)}{\sqrt{\sum_{i=1}^{k} |w_i|}} - (p_h + p_r),$$

where $\log(N/cf_i)$ is the *inverse collection frequency weight* of term t_i, N is the number of databases and cf_i is the number of databases having a positive weight value for term t_i.

ProFusion Approach. ProFusion (http://www.profusion.com) employs both training queries and real user queries for learning. In addition, ProFusion incorporates 13 preselected topic categories into the learning process (Fan & Gauch, 1999; Gauch, Wang, & Gomez, 1996). The 13 categories are "Science and Engineering," "Computer Science," "Travel," "Medical and Biotechnology," "Business and Finance," "Social and Religion," "Society, Law and Government," "Animals and Environment," "History," "Recreation and Entertainment," "Art," "Music," and "Food." For each category, a set of terms is selected to indicate the topic of the category. During the training phase, a set of training queries is identified for each category. For a given category C and a given local database D, each training query selected for C is submitted to D. From the top 10 retrieved documents, useful ones are identified by the user conducting the training. Then a score reflecting the effectiveness of D with respect to the query and the category is computed by

$$c * \frac{\sum_{i=1}^{10} N_i}{10} * \frac{R}{10},$$

where c is a constant; N_i is $1/i$ if the ith ranked document is useful, and 0 otherwise; and R is the number of useful documents in the 10 retrieved documents. This formula captures both the *rank order* of each useful document and the *precision* of the top 10 retrieved documents. Finally, the scores of database D using all training queries selected for category C is averaged to yield the *confidence factor* of D with respect to category C. At the end of the training phase, a confidence factor for each database with respect to each of the 13 categories is obtained. By using the categories and dedicated training queries, how well each local database responds to queries in different categories can be learned.

After the training is completed, the metasearch engine is ready to accept user queries. ProFusion performs database selection as follows. First, each user query q is mapped to one or more categories. Query q is mapped to category C if at least one term in the set of terms associated with C appears in q. Next, the *ranking score* of each database is computed and the databases are ranked in descending order of their ranking scores. The ranking score of a database for q is the sum of the confidence factors of the database with respect to the mapped categories. In ProFusion, only the three databases with the largest ranking scores are selected to search for each query.

ProFusion ranks retrieved documents in descending order of the product of each document's local similarity and the ranking score of the database from which the document is retrieved. Among the documents returned to the user, let d from database D be the one clicked by the user first. If the ranking algorithm were perfect, then d would be ranked at the top among all returned documents. Therefore, if d is not ranked at the top, then some adjustment should be made to fine-tune the ranking system. In ProFusion, when the first clicked document d is not ranked at the top, the ranking score of D is increased while the ranking scores of those databases whose documents are ranked higher than d are reduced. This is carried out by proportionally adjusting the confidence factors of D in mapped categories. Clearly, with this ranking score adjustment policy, document d is likely to be ranked higher if the same query is processed in the future.

Collection Fusion

After the database selector has chosen the local search engines for a given query, the next task is to determine what pages to retrieve from each selected search engine and how to merge them into a single ranked list. Different techniques to implement the document selector will be presented in Document Selection. The merging of results from multiple search engines will be covered in Result Merging.

Document Selection

A search engine typically retrieves pages in descending order of the locally computed desirabilities of the pages. Therefore, the problem of selecting what pages to retrieve from a local database can be translated into the problem of how many pages to retrieve from the database. If k pages are to be retrieved from a local database, then the k highest ranked pages will be retrieved.

A simple document selector can request that each selected search engine returns all the pages retrieved from the search engine. This approach may cause too many pages to be returned from each local system, leading to higher communication cost and more result merging effort. Another simple method for implementing a

document selector is to utilize the fact that most search engines return retrieved results in groups. Usually, only the top 10 to 20 results are returned in the first result page but the user can make additional requests for more result pages and more results. Hence, a document selector may ask each search engine to return the first few result pages. This method tends to return the same number of pages from each selected search engine. Since different search engines may contain different numbers of useful pages for a given query, retrieving the same number of pages from each search engine is likely to cause over-retrieval from less useful databases and under-retrieval from highly useful databases.

More elaborate document selection methods try to tie the number of pages to retrieve from a search engine to the ranking score (or the rank) of the search engine relative to the ranking scores (or ranks) of other search engines. This can lead to proportionally more pages to be retrieved from search engines that are ranked higher or have higher ranking scores. This type of approach is referred to as a *weighted allocation* approach in (Meng et al., 2002).

For each user query, the database selector of the metasearch engine computes a rank (i.e., 1st, 2nd, ...) and a ranking score for each local search engine. Both the rank information and the ranking score information can be used to determine the number of pages to retrieve from different local search engines. For example, in the D-WISE system (Yuwono & Lee, 1997), the ranking score information is used. Suppose for a given query q, r_i denotes the ranking score of the local database D_i, $i = 1, \ldots, k$, where k is the number of selected local databases for the query, and $\alpha = \sum_{j=1}^{k} r_j$ denotes the total ranking score for all selected local databases. D-WISE uses the ratio r_i/α to determine how many pages should be retrieved from D_i. More precisely, if m pages across these k databases are to be retrieved, then D-WISE retrieves $m * r_i/\alpha$ pages from database D_i. An example system that uses the rank information to select documents is CORI Net (Callan et al., 1995). Specifically, if m is the total number of pages to be retrieved from k selected local search engines, then

$$m * \frac{2(1+k-i)}{k(k+1)}$$

pages are retrieved from the ith ranked local database, $i = 1, \ldots, k$. Since

$$\frac{2(1+k-u)}{k(k+1)} > \frac{2(1+k-v)}{k(k+1)}$$

for $u < v$, more pages will be retrieved from the uth ranked database than from the vth ranked database. Because

$$\sum_{i=1}^{k} \frac{2(1+k-i)}{k(k+1)} = 1,$$

exactly m pages will be retrieved from the k top-ranked databases. In practice, it may be wise to retrieve slightly more than m pages from local databases in order to reduce the likelihood of missing useful pages.

It is possible to combine document selection and database selection into a single integrated process. In Database Selection, we described a method for ranking databases in descending order of the estimated similarity of the most similar document in each database for a given query. A combined database selection and document selection method for finding the m most similar pages based on these ranked databases was proposed in Yu et al. (1999). This method is sketched below. First, for some small positive integer s (e.g., s can be 2), each of the $s$top-ranked databases are searched to obtain the actual global similarity of its most similar page. This may require some locally top-ranked pages to be retrieved from each of these databases. Let *min_sim* be the minimum of these s similarities. Next, from these s databases, retrieve all pages whose actual global similarities are greater than or equal to *min_sim*. If m or more pages have been retrieved, then sort them in descending order of similarities, return the top m pages to the user, and terminate this process. Otherwise, the next top ranked database (i.e., the $(s+1)$th ranked database) is considered and its most similar page is retrieved. The actual global similarity of this page is then compared with the current *min_sim* and the minimum of these two similarities will be used as the new *min_sim*. Then retrieve from these $s+1$ databases all pages whose actual global similarities are greater than or equal to the new *min_sim*. This process is repeated until m or more pages are retrieved and the m pages with the largest similarities are returned to the user. A seeming problem with this combined method is that the same database may be searched multiple times. In practice, this problem can be avoided by retrieving and caching an appropriate number of pages when a database is searched for the first time. In this way, all subsequent "interactions" with the database would be carried out using the cached results. This method has the following property (Yu et al., 1999). If the databases containing the m desired pages are ranked higher than other databases and the similarity (or desirability) of the mth most similar (desirable) page is distinct, then all of the m desired pages will be retrieved while searching at most one database that does not contain any of the m desired pages.

Result Merging

Ideally, a metasearch engine should provide local system transparency to its users. From a user's point of view, such a transparency means that a metasearch search should behave like a regular search engine. That is, when a user submits a query, the user does not need to be aware that multiple search engines may be used to process this query, and when the user receives the search result from the metasearch engine, he/she should be hidden from the fact that the results are retrieved from multiple search engines. Result merging is a necessary task in providing the above transparency. When merging the results returned from multiple search engines into a single result, pages in the merged result should be ranked in descending order of global similarities (or global desirabilities). However, the heterogeneities that exist among local search engines and between the metasearch engine and local search engine make result merging a challenging problem. Usually, pages returned from a local search engine are ranked based on these pages' local similarities. Some local search engines make the local similarities of returned pages available to the

user (as a result, the metasearch engine can also obtain the local similarities) while other search engines do not make them available. For example, Google and AltaVista do not provide local similarities while Northern Light and FirstGov do. To make things worse, local similarities returned from different local search engines, even when made available, may be incomparable due to the use of different similarity functions and term-weighting schemes by different local search engines. Furthermore, the local similarities and the global similarity of the same page may be quite different still as the metasearch engine may use a similarity function different from those used in local systems. In fact, even when the same similarity function were used by all local systems and the metasearch engine, local and global similarities of the same page may still be very different. This is because some statistics used to compute term weights, for example the document frequency of a term, are likely to be different in different systems.

The challenge here is how to merge the pages returned from multiple local search engines into a single ranked list in a reasonable manner in the absence of local similarities and/or in the presence of incomparable similarities. An additional complication is that retrieved pages may be returned by different numbers of local search engines. For example, one page could be returned by one of the selected local search engines and another may be returned by all of them. The question is whether and how this should affect the ranking of these pages.

Note that when we say that a page is returned by a search engine, we really mean that the URL of the page is returned. One simple approach that can solve all of the above problems is to actually fetch/download all returned pages from their local servers and compute their global similarities in the metasearch engine. One metasearch engine that employs this approach for result merging is the Inquirus system (http://www.neci.nec.com/~lawrence/inquirus.html). Inquirus ranks pages returned from local search engines based on analyzing the contents of downloaded pages, and it employs a ranking formula that combines similarity and proximity matches (Lawrence & Lee Giles, 1998). In addition to being able to rank results based on desired global similarities, this approach also has some other advantages (Lawrence & Lee Giles, 1998). For example, when attempting to download pages, obsolete URLs can be discovered. This helps to remove pages with dead URLs from the final result list. In addition, downloading pages on the fly ensures that pages will be ranked based on their current contents. In contrast, similarities computed by local search engines may be based on obsolete versions of Web pages. The biggest drawback of this approach is its slow speed as fetching pages and analyzing them on the fly can be time consuming.

Most result merging methods utilize the local similarities or local ranks of returned pages to perform merging. The following cases can be identified:

Selected Databases for a Given Query Do Not Share Pages, and All Returned Pages Have Local Similarities Attached. In this case, each result page will be returned from just one search engine. Even though all returned pages have local similarities, these similarities may be normalized using different ranges by different local search engines. For example, one search engine may normalize its similarities between 0 and 1 and another between 0 and 1000. In this case, all local similarities should be renormalized based on a common range, say [0, 1], to improve the comparability of these local similarities (Dreilinger & Howe, 1997; Selberg & Etzioni, 1997).

Renormalized similarities can be further adjusted based on the usefulness of different databases for the query. Recall that when database selection is performed for a given query, the usefulness of each database is estimated and is represented as a score. The database scores can be used to adjust renormalized similarities. The idea is to give preference to pages retrieved from highly ranked databases. In CORI Net (Callan et al., 1995), the adjustment works as follows. Let s be the ranking score of local database D and $\overline{s}$ be the average of the scores of all searched databases for a given query. Then the following weight is assigned to D : $w = 1 + k * (s - \overline{s})/\overline{s}$, where k is the number of databases searched for the given query. It is easy to see from this formula that databases with higher scores will have higher weights. Let x be the renormalized similarity of page p retrieved from D. Then CORI Net computes the adjusted similarity of p by $w * x$. The result merger lists returned pages in descending order of adjusted similarities. A similar method is used in ProFusion (Gauch et al., 1996). For a given query, the adjusted similarity of a page p from a database D is the product of the renormalized similarity of p and the ranking score of D.

Selected Databases for a Given Query Do Not Share Pages, but Some Returned Pages Do Not Have Local Similarities Attached. Again, each result page will be returned by one local search engine. In general, there are two types of approaches for tackling the result-merging problem in this case. The first type uses the local rank information of returned pages directly to perform the merge. Note that in this case, local similarities that may be available for some returned pages would be ignored. The second type first converts local ranks to local similarities and then applies techniques described for the first case to perform the merge.

One simple way to use rank information only for result merging is as follows (Meng et al., 2002). First, arrange the searched databases in descending order of usefulness scores. Next, a round-robin method based on the database order and the local page rank order is used to produce an overall rank for all returned pages. Specifically, in the first round, the top-ranked page from each searched database is taken and these pages are ordered based on the database order such that the page order and the database order are consistent; if not enough pages have been obtained, the second round starts, which takes the second highest-ranked page from each searched database, orders these pages again based on the database order, and places them behind those pages selected earlier. This process is repeated until the desired number of pages is obtained.

In the D-WISE system (Yuwono & Lee, 1997), the following method for converting ranks into similarities is employed. For a given query, let r_i be the ranking score of

database D_i, r_{min} be the smallest database ranking score, r be the local rank of a page from D_i, and g be the converted similarity of the page. The conversion function is $g = 1 - (r - 1) * F_i$, where $F_i = r_{min}/(m * r_i)$ and m is the number of documents desired across all searched databases. This conversion has the following properties. First, all locally top-ranked pages have the same converted similarity, i.e., 1. Second, F_i is the difference between the converted similarities of the jth and the $(j + 1)$th ranked pages from database D_i, for any $j = 1, 2, \ldots$. Note that the distance is larger for databases with smaller ranking scores. Consequently, if the rank of a page p in a higher rank database is the same as the rank of a page p' in a lower rank database and neither p nor p' is top-ranked, then the converted similarity of p will be higher than that of p'. This property can lead to the selection of more pages from databases with higher scores into the merged result. As an example, consider two databases D_1 and D_2. Suppose $r_1 = 0.2$, $r_2 = 0.5$, and $m = 4$. Then $r_{min} = 0.2$, $F_1 = 0.25$, and $F_2 = 0.1$. Thus, the three top-ranked pages from D_1 will have converted similarities 1, 0.75, and 0.5, respectively, and the three top-ranked pages from D_2 will have converted similarities 1, 0.9, and 0.8, respectively. As a result, the merged list will contain three pages from D_2 and one page from D_1.

Selected Databases for a Given Query Share Pages. In this case, the same page may be returned by multiple local search engines. Result merging in this situation is usually carried out in two steps. In the first step, techniques discussed in the first two cases can be applied to all pages, regardless of whether they are returned by one or more search engines, to compute their similarities for merging. In the second step, for each page p returned by multiple search engines, the similarities of p due to multiple search engines are combined in a certain way to generate a final similarity for p. Many combination functions have been proposed and studied (Croft, 2000), and some of these functions have been used in metasearch engines. For example, the *max* function is used in ProFusion (Gauch et al., 1996), and the *sum* function is used in MetaCrawler (Selberg & Etzioni, 1997).

CONCLUSION

In the past decade, we have all witnessed the explosion of the Web. Up to now, the Web has become the largest digital library used by millions of people. Search engines and metasearch engines have become indispensable tools for Web users to find desired information.

While most Web users probably have used search engines and metasearch engines, few know the technologies behind these wonderful tools. This chapter has provided an overview of these technologies, from basic ideas to more advanced algorithms. As can be seen from this chapter, Web-based search technology has its roots from text retrieval techniques, but it also has many unique features. Some efforts to compare the quality of different search engines have been reported (for example, see (Hawking, Craswell, Bailey, & Griffiths, 2001)). An interesting issue is how to evaluate and compare the effectiveness of different techniques. Since most search engines employ multiple techniques, it is difficult to isolate the effect of a particular technique on effectiveness even when the effectiveness of search engines can be obtained.

Web-based search is still a pretty young discipline, and it still has a lot of room to grow. The upcoming transition of the Web from mostly HTML pages to XML pages will probably have a significant impact on Web-based search technology.

ACKNOWLEDGMENT

This work is supported in part by NSF Grants IIS-9902872, IIS-9902792, EIA-9911099, IIS-0208574, IIS-0208434 and ARO-2-5-30267.

GLOSSARY

Authority page A Web page that is linked from hub pages in a group of pages related to the same topic.

Collection fusion A technique that determines how to retrieve documents from multiple collections and merge them into a single ranked list.

Database selection The process of selecting potentially useful data sources (databases, search engines, etc.) for each user query.

Hub page A Web page with links to important (authority) Web pages all related to the same topic.

Metasearch engine A Web-based search tool that utilizes other search engines to retrieve information for its user.

PageRank A measure of Web page importance based on how Web pages are linked to each other on the Web.

Search engine A Web-based tool that retrieves potentially useful results (Web pages, products, etc.) for each user query.

Result merging The process of merging documents retrieved from multiple sources into a single ranked list.

Text retrieval A discipline that studies techniques to retrieve relevant text documents from a document collection for each query.

Web (World Wide Web) Hyperlinked documents residing on networked computers, allowing users to navigate from one document to any linked document.

CROSS REFERENCES

See *Intelligent Agents; Web Search Fundamentals; Web Site Design*.

REFERENCES

Bergman, M. (2000). The deep Web: Surfacing the hidden value. Retrieved April 25, 2002, from http://www.completeplanet.com/Tutorials/DeepWeb/index.asp

Callan, J. (2000). Distributed information retrieval. In W. Bruce Croft (Ed.), *Advances in information retrieval: Recent research from the Center for Intelligent Information Retrieval* (pp. 127–150). Dordrecht, The Netherlands: Kluwer Academic.

Callan, J., Connell, M., & Du, A. (1999). Automatic discovery of language models for text databases. In *ACM SIGMOD Conference* (pp. 479–490). New York: ACM Press.

Callan, J., Croft, W., & Harding, S. (1992). The INQUERY retrieval system. In *Third DEXA Conference, Valencia, Spain* (pp. 78–83). Wien, Austria: Springer-Verlag.

Callan, J., Lu, Z., & Croft, W. (1995). Searching distributed collections with inference networks. In *ACM SIGIR Conference, Seattle* (pp. 21–28). New York: ACM Press.

Chakrabarti, S., Dom, B., Raghavan, P., Rajagopalan, S., Gibson, D., Kleinberg, J. (1998). Automatic resource compilation by analyzing hyperlink structure and associated text. In *7th International World Wide Web Conference, Brisbane, Australia* (pp. 65–74). Amsterdam, The Netherlands: Elsevier.

Chakrabarti, S., Dom, B., Kumar, R., Raghavan, P., Rajagopalan, S., et al. (1999). Mining the Web's link structure. *IEEE Computer, 32,* 60–67.

Croft, W. (2000). Combining approaches to information retrieval. In W. Bruce Croft (Ed.), *Advances in information retrieval: Recent research from the Center for Intelligent Information Retrieval* (pp. 1–36). Dordrecht: Kluwer Academic.

Cutler, M., Deng, H., Manicaan, S., & Meng, W. (1999). A new study on using HTML structures to improve retrieval. In *Eleventh IEEE Conference on Tools with Artificial Intelligence, Chicago* (pp. 406–409). Washington, DC: IEEE Computer Society.

Dreilinger, D., & Howe, A. (1997). Experiences with selecting search engines using metasearch. *ACM Transactions on Information Systems, 15,* 195–222.

Fan, Y., & Gauch, S. (1999). Adaptive agents for information gathering from multiple, distributed information sources. In *AAAI Symposium on Intelligent Agents in Cyberspace, Stanford University* (pp. 40–46). Menlo Park, CA: AAAI Press.

Gauch, S., Wang, G., & Gomez, M. (1996). ProFusion: Intelligent fusion from multiple, distributed search engines. *Journal of Universal Computer Science, 2,* 637–649.

Gravano, L., Chang, C., Garcia-Molina, H., & Paepcke, A. (1997). Starts: Stanford proposal for Internet meta-searching. In *ACM SIGMOD Conference, Tucson, AZ* (pp. 207–218). New York: ACM Press.

Hawking, D., Craswell, N., Bailey, P., & Griffiths, K. (2001). Measuring search engine quality. *Journal of Information Retrieval, 4,* 33–59.

Hearst, M., & Pedersen, J. (1996). Reexamining the cluster hypothesis: Scatter/gather on retrieval results. In *ACM SIGIR Conference* (pp. 76–84). New York: ACM Press.

Kahle, B., & Medlar, A. (1991). *An information system for corporate users: Wide area information servers* (Tech. Rep. TMC199). Thinking Machine Corporation.

Kirsch, S. (1998). The future of Internet search: Infoseek's experiences searching the Internet. *ACM SIGIR Forum, 32,* 3–7. New York: ACM Press.

Kleinberg, J. (1998). Authoritative sources in a hyperlinked environment. In *Ninth ACM-SIAM Symposium on Discrete Algorithms* (pp. 668–677). Washington, DC: ACM–SIAM.

Koster, M. (1994). ALIWEB: Archie-like indexing in the Web. *Computer Networks and ISDN Systems, 27,* 175–182.

Lawrence, S., & Lee Giles, C. (1998). Inquirus, the NECi meta search engine. In *Seventh International World Wide Web Conference* (pp. 95–105). Amsterdam, The Netherlands: Elsevier.

Manber, U., & Bigot, P. (1997). The search broker. In *USENIX Symposium on Internet Technologies and Systems, Monterey, CA* (pp. 231–239). Berkeley, CA: USENIX.

Meng, W., Yu, C., & Liu, K. (2002). Building efficient and effective metasearch engines. *ACM Computing Surveys, 34,* 48–84.

Page, L., Brin, S., Motwani, R., & Winograd, T. (1998). *The PageRank citation ranking: Bring order to the Web* (Technical Report). Stanford, CA: Stanford University.

Pratt, W., Hearst, H., & Fagan, L. (1999). A knowledge-based approach to organizing retrieved documents. In *Sixteenth National Conference on Artificial Intelligence* (pp. 80–85). Menlo Park, CA: AAAI Press and Cambridge, MA: MIT Press.

Salton, G., & McGill, M. (1983). *Introduction to modern information retrieval.* New York: McCraw-Hill.

Selberg, E., & Etzioni, O. (1997). The MetaCrawler architecture for resource aggregation on the Web. *IEEE Expert, 12,* 8–14.

Wu, Z., Meng, W., Yu, C., & Li, Z. (2001). Towards a highly scalable and effective metasearch engine. In *Tenth World Wide Web Conference* (pp. 386–395). New York: ACM Press.

Yu, C., Meng, W., Liu, L., Wu, W., & Rishe, N. (1999). Efficient and effective metasearch for a large number of text databases. In *Eighth ACM International Conference on Information and Knowledge Management* (pp. 217–214). New York: ACM Press.

Yuwono, B., & Lee, D. (1997). Server ranking for distributed text resource systems on the Internet. In *Fifth International Conference on Database Systems for Advanced Applications* (pp. 391–400). Singapore: World Scientific.

Web Services

Akhil Sahai, *Hewlett-Packard Laboratories*
Sven Graupner, *Hewlett-Packard Laboratories*
Wooyoung Kim, *University of Illinois at Urbana-Champaign*

INTRODUCTION

There were two predominant trends in computing over the past decade—(i) a movement from monolithic software to distributed objects and components and (ii) an increasing focus on software for the Internet. Web services (or e-services) are a result of these two trends.

Web services are defined as distributed services that are identified by Uniform Resource Identifiers (URI's), whose interfaces and binding can be defined, described, and discovered by eXtensible Markup Language (XML) artifacts, and that support direct XML message-based interactions with other software applications over the Internet. Web services that perform useful tasks would often exhibit the following properties:

Discoverable—The foremost requirement for a Web service to be useful in commercial scenarios is that it be discovered by clients (humans or other Web services).

Communicable—Web services adopt a message-driven operational model where they interact with each other and perform specified operations by exchanging XML messages. The operational model is thus referred to as the Document Object Model (DOM). Some of preeminent communication patterns that are being used between Web services are synchronous, asynchronous, and transactional communication.

Conversational—Sending a document or invoking a method, and getting a reply are the basic communication primitives in Web services. A sequence of the primitives that are related to each other (thus, conversation) forms a complex interaction between Web services.

Secure and Manageable—Properties such as security, reliability, availability, and fault tolerance are critical for commercial Web services as well as manageability and quality of service.

As the Web services gain critical mass in the information technology (IT) industry as well as academia, a dominant computing paradigm of that of software as a monolithic object-oriented application is gradually giving way to software as a service accessible via the Internet.

THE GENESIS OF WEB SERVICES

Contrary to general public perception, the development of Web services followed a rather modest evolutionary path. The underpinning technologies of Web services borrow heavily from object-based distributed computing and development of the World Wide Web (Berners-Lee, 1996). In the chapter, we review related technologies that help shape the notion of Web services.

Tightly Coupled Distributed Software Architectures

The study of various aspects of distributed computing can be dated back as early as the invention of time-shared multiprocessing. Despite the early start, distributed computing remained impractical until the introduction of Object Management Group's (OMG) Common Object Request Broker Architecture (CORBA) and Microsoft's Distributed Component Object Model (DCOM), a distributed extension to the Component Object Model (COM). Both CORBA and DCOM create an illusion of a single machine over a network of (heterogeneous) computers and allow objects to invoke remote objects as if they were on the same machine, thereby vastly simplifying object sharing among applications. They do so by building their abstractions on more or less OS- and platform-independent middleware layers. In these software architectures, objects define a number of interfaces and advertise their services by registering the interfaces. Objects are assigned identifiers at the time of creation. The identifiers are used for

discovering their interfaces and their implementations. In addition, CORBA supports discovery of objects using descriptions of the services they provide. Sun Microsystems' Java Remote Method Invocation (Java RMI) provides a similar functionality, where a network of platform-neutral Java virtual machines provides the illusion of a single machine. Java RMI is a language-dependent solution, though the Java Native Interface (JNI) provides language independence to some extent.

The software architectures supported by CORBA and DCOM are said *tightly coupled* because they define their own binary message encoding, and thus objects are interoperable only with objects defined in the same software architecture; for example, CORBA objects cannot invoke methods on DCOM objects. Also, it is worth noting that security was a secondary concern in these software architectures—although some form of access control is highly desirable—partly because method-level/object-level access control is too fine-grained and incurs too much overhead, and partly because these software architectures were developed for use within the boundary of a single administrative domain, typically a local area network.

Loosely Coupled Distributed Software Architectures

Proliferation and increased accessibility of diverse intelligent devices in today's IT market have transformed the World Wide Web to a more dynamic, pervasive environment. The fundamental changes in computing landscape from a static client-server model to a dynamic peer-to-peer model encourage reasoning about interaction with these devices in terms of more abstract notion of *service* rather than a traditional notion of *object*. For example, printing can be viewed as a service that a printer provides; printing a document is to invoke the print service on a printer rather than to invoke a method on a proxy object for a printer.

Such services tend to be dispersed over a wide area, often crossing administrative boundaries, for better resource utilization. This physical distribution calls for more loosely coupled software architectures where scalable advertising and discovery are a must and low-latency, high-bandwidth interprocessor communication is highly desirable. As a direct consequence, a number of service-centric middleware developments have come to light. We note three distinctive systems from computer industry's research laboratories, namely, HP's client utility (e-Speak), Sun Microsystems' Jini, and IBM's TSpaces (here listed in the alphabetic order). These have been implemented in Java for platform independence.

Client Utility

HP's client utility is a somewhat underpublicized system that became the launching pad for HP's e-Speak (Karp, 2001). Its architecture represents one of the earlier forms of peer-to-peer system, which is suitable for Web service registration, discovery, and invocation (Kim, Graupner, & Sahai, 2002). The fundamental idea is to abstractly represent every element in computing as a uniform entity called "service (or resource)." Using the abstraction as a building block, it provides facilities for advertising and discovery, dynamic service composition, mediation and management, and capability-based fine-grain security. What distinguishes client utility most from the other systems is the fact that it makes advertisement and discovery visible to clients. Clients can describe their services using vocabularies and can specifically state what services they want to discover.

Jini

The Jini technology at Sun Microsystems is a set of protocol specifications that allows services to announce their presence and discover other services in their vicinity. It advocates a network-centric view of computing. However, it relies on the availability of multicast capability, practically limiting its applicability to services/devices connected with a local area network (such as home network). Jini exploits Java's code mobility and allows a service to export stub code which implements a communication protocol using Java RMI. Joining, advertisement, and discovery are done transparently from other services. It has been developed mainly for collaboration within a small, trusted workgroup and offers limited security and scalability supports.

TSpaces

IBM's TSpaces (TSpaces, 1999) is network middleware that aims to enable communication between applications and devices in a network of heterogeneous computers and operating systems. It is a network communication buffer with database capabilities, which extends Linda's Tuple space communication model with asynchrony. TSpaces supports hierarchical access control on the Tuple space level. Advertisement and discovery are implicit in TSpaces and provided indirectly through shared Tuple spaces.

Convergence of the Two Independent Trends

Web services are defined at the cross point of the evolution paths of service-centric computing and the World Wide Web. The idea is to provide service-centric computing by using the Internet as platform; services are delivered over the Internet (or intranet). Since its inception, the World Wide Web has strived to become a distributed, decentralized, all pervasive infrastructure where information is put out for other users to retrieve. It is this decentralized, distributed paradigm of information dissemination that upon meeting the concept of service-centric computing has led to the germination of the concept of Web services.

The Web services paradigm has caught the fancy of the research and development community. Many computer scientists and researchers from IT companies as well as universities are working together to define concepts, platforms, and standards that will determine how Web services are created, deployed, registered, discovered, and composed as well as how Web services will interact with each other.

WEB SERVICES TODAY

Web services are appearing on the Internet in the form of e-business sites and portal sites. For example,

priceline.com (http://www.priceline.com) and Expedia.com (http://www.expedia.com) act as a broker for airlines, hotels, and car rental companies. They offer through their portal sites statically composed Web services that have prenegotiated an understanding with certain airlines and hotels. These are mostly a business-to-consumer (B2C) kind of Web services. A large number of technologies and platforms have appeared and been standardized so as to enable the paradigm of Web services to support business-to-business (B2B) and B2C scenarios alike in a uniform manner. These standards enable creation and deployment, description, and discovery of Web services, as well as communication amongst them. We describe some preeminent standards below.

The Web Services Description Language (WSDL) is a standard to describe service interfaces and publish them together with services' access points (i.e., bindings) and supported interfaces. Once described in WSDL, Web services can be registered and discovered using the Universal Description, Discovery, and Integration (UDDI). After having discovered its partners, Web services use the Simple Object Access Protocol (SOAP), which is in fact an incarnation of the Remote Procedure Call (RPC) in XML, over the HyperText Transfer Protocol (HTTP) to exchange XML messages and invoke the partners' services. Though most services are implemented using platform-independent languages such as Java and C#, development and deployment platforms are also being standardized; J2EE and .NET are two well known ones. Web services and their users often expect different levels of security depending on their security requirements and assumption. The primary means for enforcing security are digital signature and strong encryption using the Public Key Infrastructure (PKI). SAML, XKMS, and XACML are some of recently proposed security standards. Also, many secure payment mechanisms have been defined. (See Figure 1).

Web Services Description

In traditional distributed software architectures, developers use an interface definition language (IDL) to define component interfaces. A component interface typically describes the operations the component supports by specifying their inputs and expected outputs. This enables developers to decouple interfaces from actual implementations. As Web services are envisaged as software accessible through the Web by other Web services and users, Web services need to be described so that their interfaces are decoupled from their implementations. WSDL serves as an IDL for Web services.

WSDL enables description of Web services independently of the message formats and network protocols used. For example, in WSDL a service is described as a set of endpoints. An endpoint is in turn a set of operations. An operation is defined in terms of messages received or sent out by the Web service:

Message—An abstract definition of data being communicated consisting of message parts.

Operation—An abstract definition of an action supported by the service. Operations are of the following types: one-way, request–response, solicit–response, and notification.

Port type—An abstract set of operations supported by one or more endpoints.

Binding—A concrete protocol and data format specification for a particular port type.

Port—A single endpoint defined as a combination of a binding and a network address.

Service—A collection of related endpoints.

As the implementation of the service changes or evolves over time, the WSDL definitions must be continuously updated and versioning the descriptions done.

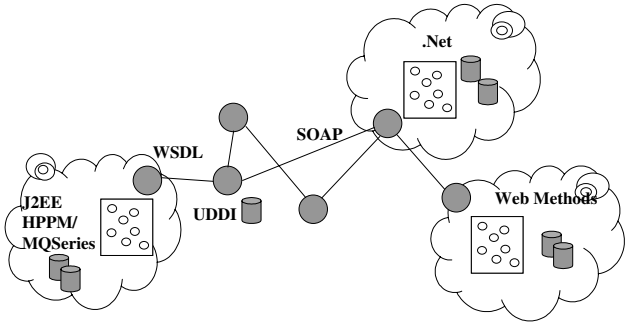


Figure 1: Web services.

Web Services Discovery

When navigating the Web for information, we use key words to find Web sites of interest through search engines. Often times, useful links in search results are mixed with a lot of unnecessary ones that need to be sifted through. Similarly, Web services need to discover compatible Web services before they undertake business with them. The need for efficient service discovery necessitates some sort of Web services clearing house with which Web services register themselves. UDDI (http://www.uddi.org) supported by Ariba, IBM, Microsoft, and HP, is an initiative to build such a Web service repository; it is now under the auspice of OASIS (http://www.oasis-open.org). These companies maintain public Web-based registries (operator sites) consistent with each other that make available information about businesses and their technical interfaces and application program interfaces (APIs). A core component of the UDDI technology is registration, an XML document defining a business and the Web services it provides. There are three parts to the registration, namely a *white page* for name, address, contact information, and other identifiers; a *yellow page* for classification of a business under standard taxonomies; and a *green page* that contains technical information about the Web services being described. UDDI also lists a set of APIs for publication and inquiry. The inquiry APIs are for browsing information in a repository (e.g., find_business, get_businessDetail). The publication APIs are for business entities to put their information on a repository.

E-marketplaces have been an important development in the business transaction arena on the Internet. They are a virtual meeting place for market participants (i.e., Web services). In addition to the basic registration

and discovery, e-marketplaces offer their participants a number of value-added services, including the following:

Enabling inter-Web service interaction after the discovery (the actual interaction may happen with or without the direct participation of the e-marketplace);

Enabling supply and demand mechanisms through traditional catalogue purchasing and request for purchase (RFP), or through more dynamic auctions and exchanges;

Enabling supply-chain management through collaborative planning and inventory handling; and

Other value-added services, such as rating, secured payment, financial handling, certification services, and notification services.

Thus, e-marketplaces can be developed as an entity that uses public UDDI registries. The e-marketplaces are categorized as vertical and horizontal depending on their target market. The vertical e-marketplaces, such as VerticalNet, GlobalNetXChange, and Retailer Market Exchange, target a specific industry sector where participants perform B2B transactions. In particular, Chemdex, E-Steel, DirectAg.com, and many more have been successful in their respective markets. By contrast, horizontal exchanges, such as eBay, are directed at a broad range of clients and businesses.

Web Services Orchestration

By specifying a set of operations in their WSDL document, Web services make visible to the external world a certain subset of internal business processes and activities. Therefore, the internal business processes must be defined and some of their activities linked to the operations before publication of the document. This in turn requires modeling a Web service's back-end business processes as well as interactions between them. On the other hand, Web services are developed to serve and utilize other Web services. This kind of interaction usually takes a form of a sequence of message exchanges and operation executions, termed *conversation*. Although conversations are described independently of the internal flows of the Web services, they result in executions of a set of backend processes. A Web service and its ensuing internal processes together form what is called a *global process*.

Intra-Web Service Modeling and Interaction

The Web Services Flow Language (WSFL) (Leymann, 2001), the Web Services Conversation Language (WSFL) (W3C, 2002), the Web Service Choreography Interface (WSCI) (BEA, 2002) and XLANG (Thatte, 2001) are some of many business process specification languages for Web services.

WSFL introduces the notion of activities and flows which are useful for describing both local business process flows and global message flows between multiple Web services. WSFL models business processes as a set of activities and links. An activity is a unit of useful work while a link connects two activities. A link can be a control link where a decision of what activity to follow is made, or a data link specifying that a certain datum flows from an activity to another. These activities may be made visible through one or more operations grouped as endpoints. As in WSDL, a set of endpoints defines a service. WSFL defines global message flows in a similar way. A global flow consists of plug links that link up operations of two service providers. Complex services involving more than two service providers are created by recursively defining plug links.

XLANG developed by Microsoft extends the XML Schema Definition Language (XSDL) to provide a mechanism for process definition and global flow coordination. The extension elements describe the behavioral aspects of a service. A behavior may span multiple operations. Action is an atomic component of a behavior definition. An action element can be an *operation,* a *delay* element, or a *raise* element. A delay element can be of type delayFor or delayUntil. delayFor and delayUntil introduce delays in execution for a process to wait for something to happen (for example, a timeout) and to wait till an absolute date-time has been reached, respectively. Raise elements are used to specify exception handling. Exceptions are handled by invoking the corresponding handler registered with a raise definition. Finally, processes combine actions in different ways: some of them are sequence, switch, while, all, pick, and empty.

Inter-Web Service Modeling and Interaction

Web services must negotiate and agree on a protocol in order to engage in a business transaction on the Web. X-EDI, ebXML, BTP, TPA-ML, cXML, and CBL have been proposed as an inter-Web service interaction protocol. We focus on ebXML as it is by far the most successful one. (See Figure 2.)

In ebXML (http://www.ebxml.org/) parties to engage in a transaction have Collaboration Protocol Profiles (CPP's) that they register at ebXML registries. A CPP contains the following:

Process Specification Layer—Details the business transactions that form the collaboration. It also specifies the order of business transactions.

Delivery Channels—Describes a party's message receiving and sending characteristics. A specification can contain more than one delivery channel.

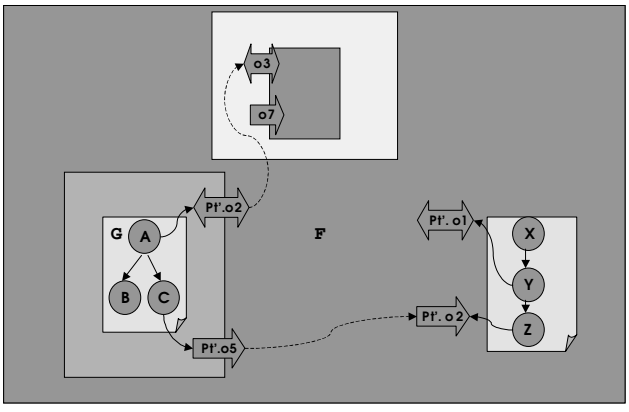


Figure 2: Intra and inter-Web service modeling and interaction.

Document Exchange Layer—Deals with processing of the business documents like digital signatures, encryption, and reliable delivery.

Transport Layer—Identifies the transport protocols to be used with the endpoint addresses, along with other properties of the transport layer. The transport protocols could be SMTP, HTTP, and FTP.

When a party discovers another party's CPP they negotiate certain agreement and form a Collaboration Protocol Agreement (CPA). The intent of the CPA is not to expose the business process internals of the parties but to make visible only the processes that are involved in interactions between the parties. Message exchange between the parties can be facilitated with the ebXML Messaging Service (ebMS). A CPA and the business process specification document it references define a *conversation* between parties. A typical conversation consists of multiple *business transactions* which in turn may involve a sequence of message exchanges for requests and replies. Although a CPA may refer to multiple business process specification documents, any conversation is allowed to involve only a single process specification document. Conceptually, the B2B servers of parties involved are responsible for managing CPAs and for keeping track of the conversations. They also interface the operations defined in a CPA with the corresponding internal business processes.

Web Services Platforms

Web services platforms are the technologies, means, and methods available to build and operate Web services. Platforms have been developed and changed over the course of time. A classification into four generations of platform technology should help to structure the space:

First Generation: HTML and CGI—Characterized by Web servers, static HTML pages, HTML FORMS for simple dialogs, and the Common Gateway Interface (CGI) to connect Web servers to application programs, mostly Perl or Shell scripts. (See Figure 3.)

Second Generation: Java—Server-side dynamic generation of HTML pages and user session support; the Java servlet interface became popular for connecting to application programs.

Third Generation: Application server as Richer development and run-time environments—J2EE as foundation for application servers that later evolved towards the fourth generation.

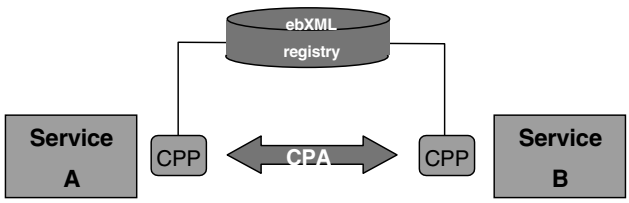


Figure 3: ebXML service-to-service interaction.

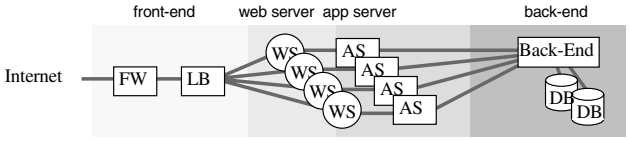


Figure 4: Basic four-tier architecture for Web services.

Fourth Generation: Web services—Characterized by the introduction of XML and WSDL interfaces for Web services with SOAP-based messaging. A global service infrastructure for service registration and discovery emerged: UDDI. Dynamic Web services aggregation—Characterized by flow systems, business negotiations, agent technology, etc.

Technically, Web services have been built according to a pattern of an *n*-tier architecture that consists of a front-end tier, firewall (FW), load balancer (LB), a Web-server tier (WS), an application (server) (AS) tier, and a back-end tier for persistent data, or the database tier (DB). (See Figure 4.)

First Generation: HTML and CGI

The emergence of the World Wide Web facilitated the easy access and decent appearance of linked HTML mark-up pages in a user's browser. In the early days, it was mostly static HTML content. Passive information services that provided users with the only capability of navigating though static pages could be built. However, HTML supported from the very beginning FORMS that allowed users to enter text or select from multiple-choice menus. FORMS were treated specially by Web servers. They were passed onto CGI, behind which small applications, mostly Perl or Shell scripts, could read the user's input, perform respective actions, and return a HTML page that could then be displayed in the user's browser. This primitive mechanism enabled a first generation of services on the Web beyond pure navigation through static contents.

Second Generation: Java

With the growth of the Web and the desire for richer services such as online shopping and booking, the initial means to build Web services quickly became too primitive. Java applets also brought graphical interactiveness to the browser side. Java appeared as the language of choice for Web services. Servlets provided a better interface between the Web server and the application. Technology to support dynamic generation of HTML pages at the server side was introduced: JSP (Java Server Pages) by Sun Microsystems, ASP (Active Server Pages) by Microsoft, or PHP pages in the Linux world enabled separation of presentation, the appearance of pages in browsers, from content data. Templates and content were then merged on the fly at the server in order to generate the final page returned to the browser. Since user identification was critical for business services, user log-in and user sessions were introduced. Applications were becoming more complex, and it turned out that there was a significant overlap in common functions needed for many services such as session support, connectivity to persistent databases, and security functions.

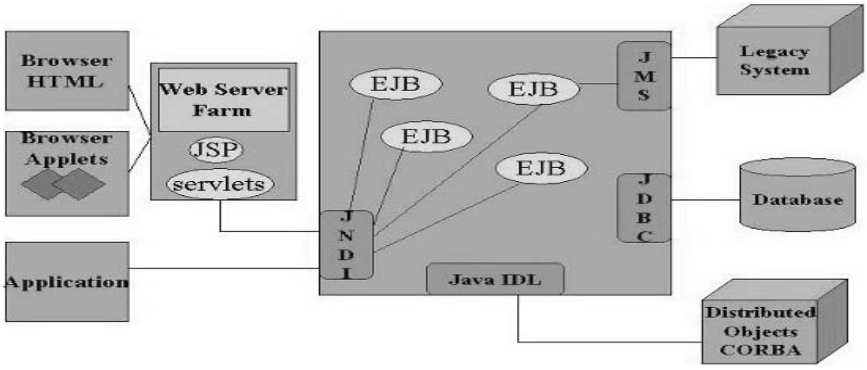


Figure 5: The J2EE platform.

Third Generation: Application Server

The observation that many functions were shared and common among Web services drove the development toward richer development environments based on the Java language and Java libraries. A cornerstone of these environments became J2EE (Java 2 Platform, Enterprise Edition), which is a Java platform designed for enterprise-scale computing. Sun Microsystems (together with industry partners such as IBM) designed J2EE (Figure 5) to simplify application development for Web services by decreasing the need for programming through reusable modular components and by providing standard functions such as session support and database connectivity.

J2EE primarily manifests in a set of libraries used by application programs performing the various functions. Web service developers still had to assemble all the pieces, link them together, connect them to the Web server, and manage the various configurations. This led to the emergence of software packages that could be deployed easier on a variety of machines. These packages later became application servers. They significantly reduced the amount of configuration work during service deployment such that service developers could spend more time on business logic and the actual function of the service. Most application server are based on J2EE technology. Examples are IBM's WebSphere suite, BEA's WebLogic environment, the Sun ONE Application Framework, and Oracle's 9*i* application server. (See Figure 5.)

Fourth Generation: Web Services

Prior generations of Web services mostly focused on end-users, people accessing services from Web browsers. However, accessing services from services other than browsers turned out to be difficult. This circumstance has prevented the occurrence of Web service aggregation for a long time. Web service aggregation meant that users would only have to contact one Web service, and this service then would resolve the user's requests with further requests to other Web services.

HTML is a language defined for rendering and presenting content in Web browsers. It does not allow per se separating content from presentation information. With the advent of XML, XML became the language of choice for Web services for providing interfaces that could not only be accessed by users through Web browsers but also by other services. XML is now pervasively being used in Web services messaging (mainly using SOAP) and for Web service interface descriptions (WSDL). In regard to platforms, XML enhancements were added to J2EE and application servers. The introduction of XML is the major differentiator between Web services platforms of the third and the fourth generation in this classification.

A major step toward the service-to-service integration was the introduction of the UDDI service (see the above section Web Services Discovery).

Three major platforms for further Web services interaction and integration are: Sun Microsystems' Sun ONE (Open Net Environment), IBM WebSphere, and Microsoft's .NET.

Sun ONE—Sun's standards-based software architecture and platform for building and deploying services on demand. Sun ONE's architecture is built around existing business assets: Data, applications, reports, and transactions, referred to as the DART model. Major standards are supported: XML, SOAP, J2EE, UDDI, LDAP, and ebXML. The architecture is composed of several product lines: the iPlanet Application Framework (JATO), Sun's J2EE application framework for enterprise Web services development, application server, portal server, integration server, directory server, e-commerce components, the Solaris Operating Environment, and development tools.

IBM WebSphere—IBM's platform to build, deploy, and integrate your e-business, including components such as foundation and tools, reach and user experience, business integration, and transaction servers and tools.

Microsoft .NET—Microsoft's .NET platform for providing lead technology for future distributed applications inherently seen as Web services. With Microsoft .NET, Web services' application code is built in discrete units, XML Web services, which handle a specified set of tasks. Because standard interfaces based on XML simplify communication among software, XML Web services can be linked together into highly specific applications and experiences. The vision is that the best XML Web services from any provider around the globe can be used to create a needed solution quickly and easily.

Microsoft will provide a core set of XML Web services, called Microsoft .NET My Services, to provide functions such as user identification and calendar access.

Security and Web Services

Due to their public nature, security is vital for Web services. Security attacks can be classified as threats of information disclosure, unauthorized alteration of data, denial of use, misuse or abuse of services, and, more rarely considered, repudiation of access. Since Web services link networks together with businesses, further attacks, such as masquerading, stealing or duplicating identity and conducting business under false identity, or accessing or transferring funds from or to unauthorized accounts, need to be considered.

Security is vital for establishing the legal basis for businesses done over the Web. Identification and authentication of business partners are the basic security requirements. Others include integrity and authenticity of electronic documents. Electronic contracts must have the same binding legal status as conventional contracts. Refusal and repudiation of electronic contracts must be provable in order to be legally valid. Finally, payment and transferring funds between accounts must be safe and secure.

Security architectures in networks are typically composed of several layers:

Secure data communication—IPsec (Internet Protocol Security), SSL (Secure Socket Layer), TLS (Transport Layer Security);

Secured networks—VPNs (Virtual Private Networks);

Authenticity of electronic documents and issuing individuals—digital signatures;

Secure and authenticated access—digital certificates;

Secure authentication and certification—PKI (Public Key Infrastructure); and

Single sign-on and digital passports.

Single Sign-On and Digital Passports

Digital passport emerged from the desire to provide an individual's identity information from a trusted and secure centralized place rather then repeatedly establishing this information with each collaborating partner and maintaining separate access credentials for each pair of collaborations. Individuals only need one such credential, the passport, in order to provide collaborating partners with certain parts of an individual's identity information. This consolidates the need for maintaining separate identities with different partners into a single identification mechanism. Digital passports provide an authenticated access to a centralized place where individuals have registered their identity information such as phone numbers, social security numbers, addresses, credit records, and payment information. Participating individuals, both people and businesses, will access the same authenticated information assuming trust to the authority providing the passport service. Two initiatives have emerged: Microsoft's .NET Passport and the Liberty Alliance Project, initiated by Sun Microsystems.

Microsoft .NET Passport (Microsoft .NET, 2002) is a single sign-on mechanism for users on the Internet. Instead of creating separate accounts and passwords with every e-commerce site, users only need to authenticate with a single Passport server. Then, through a series of authentications and encrypted cookie certificates, the user is able to purchase items at any participating e-commerce site without verifying the user's identity again. .NET Passport is an online service that enables use of an e-mail address and a single (Passport server) password to securely sign in to any .NET Passport participating Web site or service. It allows users to easily move among participating sites without the need to verify their identity again. The Microsoft .NET Passport had initially been planned for signing into Microsoft's own services. Expanding it toward broader use in the Web has been seen as critical. This concern gave reason for the Liberty Alliance Project initiative that is now widely supported in industry and public.

The Liberty Alliance Project (Liberty Alliance Project, 2002) is an organization being formed to create an open, federated, single sign-on identity solution for the digital economy via any device connected to the Internet. Membership is open to all commercial and noncommercial organizations. The Alliance has three main objectives:

1. To enable consumers and businesses to maintain personal information securely.
2. To provide a universal, open standard for single sign-on with decentralized authentication and open authorization from multiple providers.
3. To provide an open standard for network identity spanning all network-connected devices.

With the emergence of Web services, specific security technology is emerging. Two major security technology classes are Java-based security technology and XML-based security technology.

Both classes basically provide mappings of security technologies, such as authentication and authorization, encryption, and signatures, into respective environments.

Java-Based Security Technology for Web Services

Java-based security technology is primarily available through the Java 2 SDK and J2EE environments in the form of sets of libraries:

Encryption—JSSE (Java Secure Socket Extension); the JCE (Java Cryptography Extension) provides a framework and implementations for encryption, key generation and key agreement, and Message Authentication Code (MAC) algorithms. Support for encryption includes symmetric, asymmetric, block, and stream ciphers. The software also supports secure streams and sealed objects.

Secure messaging—Java GSS-API is used for securely exchanging messages between communicating applications. The Java GSS-API contains the Java bindings for the Generic Security Services Application Program Interface (GSS-API) defined in RFC 2853. GSS-API

offers application programmers uniform access to security services atop a variety of underlying security mechanisms, including Kerberos.

Authentication and Authorization—JAAS (Java Authentication and Authorization Service) for authentication of users, to reliably and securely determine who is currently executing Java code, and for authorization of users to ensure they have the access rights (permissions) required to do security-sensitive operations.

Certification—Java Certification Path API.

X.509 Certificates and Certificate Revocation Lists (CRLs) and Security Managers.

These libraries are available for use when Web services are built using Java. They are usually used when building individual Web services with application servers.

For Web services interaction, XML technology eliminates the tied binding to Java. Consequently, a similar set of XML-based security technologies enabling cross-service interactions is emerging.

XML-Based Security Technology for Web Services

The Organization for the Advancement of Structured Information Standards (OASIS) merges security into Web services at a higher level than the common Internet security mechanisms and practices described above. Proposals are primarily directed toward providing XML specifications for documents and protocols suitable for cross-organizational Web services interactions. XML-based security technology can be classified into the following:

XML Document-Level Security—encryption and digitally signing XML documents;

Protocol-Level Security for XML Document Exchanges—exchanging XML documents for authentication and authorization of peers; and

XML-Based Security Frameworks—infrastructures for establishing secure relationships among parties.

XML Document-Level Security: Encryption and Signature. The (preliminary) XML encryption specification (Reagle, 2000) details requirements on how to digitally encrypt a Web resource in general, and an XML document in particular. XML encryption can be applied to a part of or complete XML document. The granularity of encryption can be reduced to an element, attributes, or text content. Encryption can be recursive. The specification does not address confidence or trust relationships and key establishment. The specification addresses both key-encrypting-keys and data keys. The specification will not address the expression of access control policies associated with portions of the XML document. This will be addressed by XACML.

XML signature defines the XML schema and processing rules for creating and representing digital signatures in any digital content (data object), including XML. An XML signature may be applied to the content of one or more documents. Enveloped or enveloping signatures are over data within the same XML document as the signature; detached signatures are over data external to the signature element. More specifically, this specification defines an XML signature element type and an XML signature application; conformance requirements for each are specified by way of schema definitions and prose respectively. This specification also includes other useful types that identify methods for referencing collections of resources, algorithms, and keying and management information.

The XML Signature (Bartel, Boyer, Fox, LaMacchia, & Simon, 2002) is a method of associating a key with referenced data (octets); it does not normatively specify how keys are associated with persons or institutions, nor the meaning of the data being referenced and signed. Consequently, while this specification is an important component of secure XML applications, it itself is not sufficient to address all application security/trust concerns, particularly with respect to using signed XML (or other data formats) as a basis of human-to-human communication and agreement. Such an application must specify additional key, algorithm, processing, and rendering requirements. The SOAP Digital Signature Extensions defines how specifically SOAP messages can be digitally signed.

Protocol-Level Security for XML Document Exchanges. Protocol-level security defines document exchanges with the purpose of establishing secure relationships among parties, typically providing well-defined interfaces and XML bindings to an existing public key infrastructure. Protocol-level security can be built upon the document-level security.

The XML Key Management Specification (Ford et al., 2001) defines protocols for validating and registering public keys, suitable for use in conjunction with the proposed standard for XML signature developed by the World Wide Web Consortium (W3C) and the Internet Engineering Task Force (IETF) and an anticipated companion standard for XML encryption. The XML Key Management Specification (XKMS) comprises two parts: the XML Key Information Service Specification (X-KISS) and the XML Key Registration Service Specification (X-KRSS).

The X-KISS specification defines a protocol for a trust service that resolves public key information contained in XML-SIG document elements. The X-KISS protocol allows a client of such a service to delegate part or all of the tasks required to process <ds:KeyInfo> elements embedded in a document. A key objective of the protocol design is to minimize the complexity of application implementations by allowing them to become clients and thereby shielded from the complexity and syntax of the underlying Public Key Infrastructure (OASIS PKI Member Section, 2002) used to establish trust relationships-based specifications such as X.509/PKIX, or SPKI (Simple Public Key Infrastructure, 1999).

The X-KRSS specification defines a protocol for a web service that accepts registration of public key information. Once registered, the public key may be used in conjunction with other web services including X-KISS.

XML-Based Security Frameworks. XML-based security frameworks go one step further than the above.

The Security Assertion Markup Language (SAML), developed under the guidance of OASIS (OASIS, 2002), is an XML-based framework for exchanging security information with established, SAML-compliant security services. This security information is expressed in the form of assertions about subjects, where a subject is an entity (either human or program) that has an identity in some security domain. A typical example of a subject is a person, identified by his or her e-mail address in a particular Internet DNS domain.

Assertions can convey information about authentication acts performed by subjects, attributes of subjects, and authorization decisions about whether subjects are allowed to access certain resources. Assertions are represented as XML constructs and have a nested structure, whereby a single assertion might contain several different internal statements about authentication, authorization, and attributes. Assertions containing authentication statements merely describe acts of authentication that happened previously.

Assertions are issued by SAML authorities, namely, authentication authorities, attribute authorities, and policy decision points. SAML defines a protocol by which relying parties can request assertions from SAML authorities and get a response from them. This protocol, consisting of XML-based request-and-response message formats, can be bound to many different underlying communications and transport protocols. Currently it defines only one binding, namely SOAP over HTTP.

SAML authorities can use various sources of information, such as external policy stores and assertions that were received as input in requests, in creating their responses. Thus, while clients always consume assertions, SAML authorities can be both producers and consumers of assertions.

Payment Systems for Web Services

Effective payment systems are a prerequisite for business with Web services. This section introduces and classifies different approaches for payment systems that have been developed over the passed years. However, payments in the Internet are mostly conducted through the existing payment infrastructure that was developed before the Internet became pervasive. End-consumer retail business on the Internet primarily relies on credit card transactions. Other traditional payment methods are offered as well: personal checks, money orders, or invoice billing. In the business-to-business segment, traditional invoice billing is still the major payment method. An overview is given in (Weber, 1998). W3C has adopted payment standards (Micropayment Overview, 2002).

Payments by Credit Cards

The reason why credit card payments are well accepted is that credit card providers act as intermediaries between payers and recipients of payments (payees). They do also guarantee payments up to a limit (important to the payee), and they carry the risk of misuse. All parties must register accounts before transfers can be conducted. Another important service is the verification of creditability of a person or a business before opening an account.

SET—The Secure Electronic Transaction Standard

SET (Secure Electronic Transaction, 2002) is an open technical standard for the commerce industry initially developed by two major credit card providers, Visa and MasterCard, as a way to facilitate secure payment card transactions over the Internet. Digital certificates (Digital Certificates, 1988) create a trust chain throughout the transaction, verifying cardholders' and merchants' identity. SET is a system for ensuring the security of financial transactions of credit card providers or bank accounts. Its main objective is to provide a higher security standard for credit card payments on the Internet. A major enhancement compared to traditional credit card payments is that neither credit card credentials nor payers' identity are revealed to merchants. With SET, a user is given an electronic wallet (digital certificate). A transaction is conducted and verified using a combination of digital certificates and digital signatures among the purchaser, a merchant, and the purchaser's bank in a way that ensures privacy and confidentiality.

Not all payments required by Web services can be conducted through credit card transactions. First, credit card transactions are typically directed from an end-customer, a person, to a business that can receive such payments. Second, the amounts transferred through a credit card transaction are limited to a range between currency equivalents of > \$0.10 up to several thousand dollars depending on an individual's credit limits. Micropayments <\$0.10, as well as macropayments> \$10,000, are typically not provided. The lower payment bound is also caused by the cost per transaction model credit card providers use. Third, payments among persons, as for instance required for auctions among people or for buying and selling used goods, cannot be conducted through credit card accounts. Traditional payment methods are used here: personal checks, money orders, or cash settlement. Fourth, only individuals with registered accounts can participate in credit card payments. Individuals that do not qualify are excluded. This restriction is also a major barrier for Web service business in developing countries.

Micropayments

The purpose of micropayments is primarily for "pay-per-use" models where the usage is measured and immediately charged to customers in very small amounts. Transaction costs for micropayment systems need to be significantly lower, and the number of transactions may be significantly higher than that of credit card payments. Accurate, fine-grained charging is enabled. These are the two major differentiators of micropayment systems. W3C proposes the Common Markup for Micropayment "per-fee-links."

Micropayments involve a buyer or customer, a vendor or merchant, and potentially one or more additional parties that keep accounts in order to aggregate micro payments for final charge. These mediators are called brokers (in Millicent), billing servers (in IBM MicroPayments), or intermediaries (in France Telecom Micropayments), to name a few.

Millicent. One micropayment system is Millicent (Glassman, 2000). The MilliCent Microcommerce

Network provides new pay-per-click/earn-per-click functionality for Internet users. It allows buying and selling digital products costing from 1/10th of a cent to up to $10.00 or more. MilliCent can be used by Web services to build any number of parallel revenue streams through the simultaneous use of pay-per-click purchases, subscriptions, and advertising. It can also be used to make direct monetary payments to users. MilliCent is optimized for buying and selling digital products or services over the Internet such as articles, newsletters, real-time data, streaming audio, electronic postage, video streams, maps, financial data, multimedia objects, interactive games, software, and hyperlinks to other sites.

NetBill. NetBill is a Carnegie Mellon University Internet billing server project, which is used as a payment method for buying information goods and services via the Internet. It aims at secure payment for and delivery of information goods, e.g., library services, journal articles, and CPU cycles. The NetBill system charges for transactions and requires customers to have a prepaid NetBill account from which all payments are deducted. The NetBill payment system uses both symmetric key and public key cryptography. It relies on Kerberos for authentication. An account server, called NetBill server, maintains accounts for both customers and merchants. NetBill acts as an aggregator to combine many small transactions into larger conventional transactions, thus amortizing conventional overhead fees. Customers and merchants have to trust the NetBill server.

Digital Money and Digital Coins

In contrast to account-based payment systems, such as credit card-based systems, where amounts are transferred between accounts inside or between credit card or bank providers, digital money represents a value amount flowing from a payer to a payee across the network. Establishing accounts with providers before services can actually be used is unnecessary. Advantages are the same as for cash money: no mutual accounts need to be established before a payment can be conducted. No mutual authentication is needed for improving convenience for both parties. In addition, as with cash money, the payer does not need to reveal any identity credentials to the payee or someone else. Payments are anonymous and nontraceable. A major hurdle for this approach is the prevention of duplication and forging of digital money since no physical security marks such as watermarks can be applied to digitized bit strings.

The basic idea behind digital money is that a consumer purchases "digital coins" from an issuer using a regular payment method such as a credit card. The issuer generates an account for that customer and deposits the amount into it. It then hands out a set of digital coins to the customer that he or she can use for payments. For a payment, the customer transfers coins to the merchant or service provider. The provider then transfers coins to the issuer and deposits them into his account. The merchant, however, may also use these coins to pay its suppliers. Digital coins will thus flow among participants similarly like cash money flows among people.

The following requirements need to be met by digital money systems:

digital money must be protected from duplication or forging; and

digital money should neither contain nor reveal identity credentials of any involved party in order to be anonymous.

The first requirement is achieved by not actually representing an amount by a digital coin, but rather a reference to an allocated amount in the possessor's account with the issuer. When digital coins are copied, the reference is copied, not the amount itself. However, the first individual redeeming a coin with the issuer will receive the amount. Identity at redemption cannot be verified since digital coins do not carry identifying credentials of the possessor. The only term the issuer can verify is whether or not a coin has already been redeemed. By thus, theft of digital money is possible, and parties have an interest in keeping their coins protected.

Achieving complete anonymity between an issuer and subsequent receivers of digital money is a key characteristic of digital money. It is basically achieved by blinded signatures (Chaum, 1985) that guarantee to uniquely assign coins with allocated amounts within the issuer's account system and without revealing any identification information of the holder of that account.

E-cash. E-cash (CrytoLogic Ecash FAQ, 2002) stands for "electronic cash," a system developed by DigiCash that underwent field tests in the late 1990s. E-cash is a legal form of computer-generated currency. This currency can be securely purchased with conventional means: credit cards, checks, money orders, or wire transfers.

MicroMint. MicroMint is a proposal by Rivest and Shamir about coins that can only efficiently be produced in very large quantities and are hard to produce in small quantities. The validity of a coin is easily checked. MicroMint is optimized for unrelated low-value payments. It uses no public key operations. However, the scheme is very complex and would require a lot of initial and operational efforts. Therefore, it is unlikely that it ever will gain any practical importance.

A broker will issue new coins at the beginning of a period and will revoke those of the prior period. Coins consist of multiple hash collisions, i.e., different values that all hash to the same value. The broker mints coins by computing such hash collisions. For that process many computations are required, but more and more hash collisions are detected with continued computation. The broker sells these MicroMint coins in batches to customers. Unused coins can be returned to the broker at the end of a period, e.g., a month. Customers render MicroMint coins as payment to merchants.

THE FUTURE OF WEB SERVICES

In future we will see the unleashing of a Web services phenomenon. This will involve the fulfillment of dynamic Web service composition and orchestration vision, the appearance of personalized Web services, concepts of Web

service management, and the development of Web service infrastructure as a reusable, reconfigurable, self-healing, self-managing, large-scale system.

Dynamic Web Services Composition and Orchestration

The vision of Web services intelligently interacting with one another and performing useful tasks automatically and seamlessly remains to become reality. Major milestones have been achieved: XML as a syntactic framework and data representation language for Web services interaction; the Web infrastructure itself providing ubiquitous access to Web services; the emergence of global registration and discovery services; and the technology to support the creation and maintenance of Web services, just to name a few. However, major pieces such as the formalization and description of service semantic are yet to be developed. The effort of creating a semantic Web (Semantic Web, 2001) is an extension of the current Web in which information is given well-defined meaning, better enabling computers and people to work in cooperation. Ontologies define the structure, relationships, and meaning of terms appearing in service descriptions. The semantic Web vision is that these ontologies can be registered, discovered, and used for reasoning about Web service selection before undertaking business. Languages like DAML+OIL (DAML, 2001) have been developed in this context.

In addition, sending a document or invoking a method and getting a reply are the basic communication primitives. However, complex interactions between Web services will involve multiple steps of communication that are related to each other. A conversation definition is a sequencing of document exchanges (method invocations in the network object model) that together accomplish some business functionality. In addition to agreeing upon vocabularies and document formats, conversational Web services also agree upon conversation definitions before communicating with each other. A conversation definition consists of descriptions of interactions and transitions. Interactions define the atomic units of information interchange between Web services. Essentially, each service describes each interaction in terms of the documents that it will accept as input or will produce as output. The interactions are the building blocks of the conversation definition. Transitions specify the ordering amongst the interactions. Web services need to introspect other Web services and obtain each other's descriptions before they start communicating and collaborating (Banerji et al., 2002).

RosettaNet (RosettaNet, 2002) is a nonprofit consortium of major information technology, electronic components, and semiconductor manufacturing companies working to create and implement industry-wide, open e-business process standards, particularly targeting business-to-business market places, workflow, and supply-chain management solutions. These standards form a common e-business language, aligning processes between supply-chain partners on a global basis. Several examples exist. The centerpiece of the RosettaNet model is the partner interface process (PIP). The PIP defines the activities, decisions, and interactions that each e-business trading participant is responsible for. Although the RosettaNet model has been in development, it will be a while until Web services start using them to undertake business on the Web.

Once these hurdles are overcome, the basis and platform for true Web services that will enable agent technologies merging into Web services to provide the envisioned dynamic Web service aggregation on demand according to users' specifications will emerge.

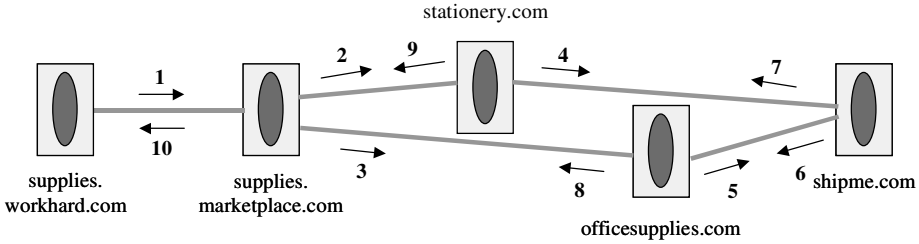
Personalized Web Services

As Web service technology evolves, we anticipate that they will become increasingly sophisticated, and that the challenges the Web service community will face will also evolve to meet their new capabilities. One of the most important of these challenges is the question of what it means to personalize Web services. Personalization can be achieved by using user profiles, i.e., monitoring user behavior, devices, and context to customize Web services (Kuno & Sahai, 2002) for achieving metrics like quality of experience (QoE) (van Moorsel, 2001). This would involve providing and meeting guarantees of service performance on the user's side. Personalization could also result in the creation of third-party rating agencies that will register user experiences, which could be informative for other first-time users. These rating mechanisms already exist in an ad hoc manner, e.g., eBay and Amazon allow users to rate sellers and commodities (books), respectively. Salcentral.com and bizrate.com are third-party rating agencies that rate businesses. These services could be also developed as extended UDDI services. These mechanisms will also render Web services more "customer-friendly."

End-to-End Web Service Interactions

Web services are federated in nature as they interact across management domains and enterprise networks. Their implementations can be vastly different in nature. When two Web services connect to each other, they must agree on a document exchange protocol and the appropriate document formats (Austin, Barbir, & Garg 2002). From then on they can interoperate with each other, exchanging documents. SOAP defines a common layer for document exchange. Services can define their own service-specific protocol on top of SOAP. Often, these Web service transactions will span multiple Web services. A request originating at a particular Web service can lead to transactions on a set of Web services. For example, a purchase order transaction that begins when an employee orders supplies and ends when he or she receives a confirmation could result in 10 messages being exchanged between various services as shown in Figure 6.

The exchange of messages between Web services could be asynchronous. Services sending a request message need not be blocked waiting for a response message. In some cases, all the participating services are like peers, in which case there is no notion of a request or a response. Some of the message flow patterns that result from this asynchrony are shown in Figure 7. The first example in Figure 7 shows a single request resulting in multiple responses. The second example shows a broker-scenario,



1. purchase order
2. part of purchase order
3. the other part of the purchase order
4. shipping request
5. shipping request
6. shipping confirmation
7. shipping confirmation
8. order confirmation
9. order confirmation
10. purchase order confirmation

Figure 6: SOAP messages exchanged between Web services.

in which a request is sent to a broker but responses are received directly from a set of suppliers.

These Web services also interact with a complex web of business processes at their back-ends. Some of these business processes are exposed as Web service operations. A business process comprises a sequence of activities and links as defined by WSFL and XLANG. These business processes must be managed so as to manage Web service interactions. Management of Web services thus is a challenging task because of their heterogeneity, asynchrony, and federation. Managing Web services involves managing business transactions by correlation of messages across enterprises (Sahai, Machiraju, & Wurster, 2001) and managing the business processes.

Also, in order to manage business on the Web, users will need to specify, agree, and monitor service level agreements (SLAs) with each other. Thus, Web services will invariably have a large number of SLAs. As less human intervention is more desirable, the large number of SLAs would necessitate automating the process as much as possible (Sahai, Machiraju, Sayal, Jin, & Casati, 2002).

Web service to Web service interaction management can also be done through mediation (Machiraju, Sahai, & van Moorsel, 2002). Web service networks' vision is to mediate Web service interactions, so as to make it secure, manageable, and reliable. Such networks enable versioning management, reliable messaging, and monitoring of message flows (e.g., Flamenco Networks, GrandCentral, Transact Plus, Talking Blocks).

Future Web Services Infrastructures

Deployment and operational costs are determinants in the balance sheets for Web service providers. Web service providers are optimizing their IT infrastructures to allow faster provisioning of new services and more reliable operation. Platforms and management solutions that reduce Web services' deployment and operational costs are emerging. Those platforms support the deployment of Web services (installation and configuration of software and content data), the virtual wiring of machines into application environments independently of the physical wiring in a data center. They allow rearrangements of Web services' applications among machines, the dynamic sizing of service capacities according to fluctuations in demands, and the isolation of service environments hosted in the same data center.

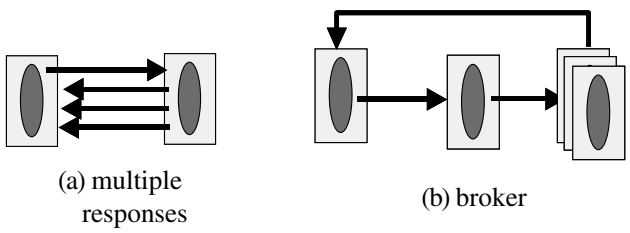


Figure 7: Asynchronous message patterns between Web services.

HP's Utility Data Center (HP Utility Data Center, 2001) is such a platform. The HP Utility Data Center with its Utility Controller Software creates and runs virtual IT environments as a highly automated service optimizing asset utilization and reducing staffing loads. Resource virtualization is invisible to applications, sitting underneath the abstractions of operating systems.

Two types of resources are virtualized:

Virtualized network resources, permitting the rewiring of servers and related assets to create entire virtual IT environments; and

Virtualized storage resources, for secure, effective storage partitioning, and with disk images containing persistent states of application environments such as file systems, bootable operating system images, and application software.

Figure 8 shows the basic building blocks of such a utility data center with two fabrics for network virtualization and storage virtualization.

The storage virtualization fabric with the storage area network attaches storage elements (disks) to processing elements (machines). The network virtualization fabric then allows linking processing elements together in a virtual LAN.

Two major benefits for Web services management can be achieved on top of the infrastructure:

Automated Web services deployment—By entirely maintaining persistent Web services' states in the storage system and conducting programmatic control over

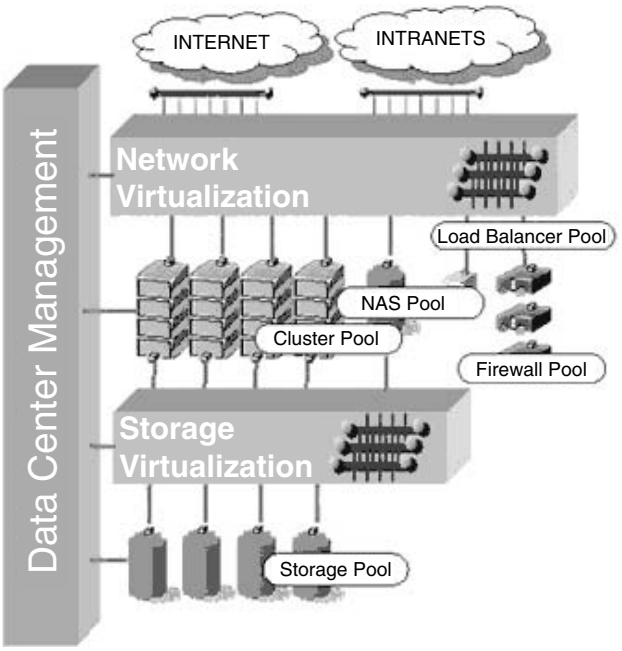


Figure 8: Architecture of a utility data center Web services platform infrastructure.

storage containing deployed service arrangements; and

Dynamic capacity sizing of Web services—By the ability to automatically launch additional service instances absorbing additional load occurring to the service. Service instances are launched by first allocating spare machines from the pool maintained in the data center, wiring them into the specific environment of the Web service, attaching appropriate storage to those machines, and launching the applications obtained from that storage. Web server farms are a good example for such a "breathing" (meaning dynamically adjustable) configuration (Andrzejak, Graupner, Kotov, & Trinks, 2002; Graupner, Kotov, & Trinks, 2002).

IBM's Autonomic Computing vision is to provide for self-managing systems. The intent is to create systems that respond to capacity demands and system glitches without human intervention. These systems intend to be self-configuring, self-healing, self-protecting, and self-optimizing (IBM Autonomic Computing, 2002).

CONCLUSION

The Web services paradigm has evolved substantially because of concerted efforts by the software community. The genesis of Web services can be traced back to projects like e-speak, Jini, and TSpaces. Although progress has been made in Web service standardization, the full potential of Web services remains unrealized. The future will see the realization of Web services as a means of doing business on the Web, the vision of dynamic composition of Web services, personalized Web services, end-to-end management of Web service interactions, and a dynamically reusable service infrastructure that will adapt to variations in resource consumption.

GLOSSARY

Business process execution language for Web services (BPEL4WS) A standard business process description language that combines features from WSFL and XLANG.

Composition Creating composite Web services when Web services outsource their functionalities to other Web services.

Conversation A set of message exchanges that can be logically grouped together.

Description Describing Web services in terms of the operations and messages they support, so that they can be registered and discovered at UDDI operator sites or by using WS-Inspection.

End-to-end management Protocol required to track and manage Web service composition leading to a transaction being subdivided amongst multiple Web services.

Orchestration Web service to Web service interaction that leads to the coupling of internal business processes.

Personalization Personalizing or customizing Web services to user/client profiles and requirements.

Platform One or more execution engines over which a Web service implementation is executed.

Service level agreement (SLA) An agreement that specifies quality-of-service guarantees between parties.

Simple object access protocol (SOAP) A standard for messaging between Web services.

Web service conversation language (WSCL) A language to describe Web service conversations.

Web services flow language (WSFL) A language to describe business processes.

CROSS REFERENCES

See *Client/Server Computing; Common Gateway Interface (CGI) Scripts; Electronic Payment; Java; Perl; Personalization and Customization Technologies; Secure Electronic Transmissions (SET)*.

REFERENCES

Andrzejak, A., Graupner, S., Kotov, V., & Trinks, H. (2002). Self-organizing control in planetary-scale computing. In *IEEE International Symposium on Cluster Computing and the Grid (CCGrid), 2nd Workshop on Agent-based Cluster and Grid Computing (ACGC)*. New York: IEEE.

Austin, D., Barbir, A., & Garg, S. (2002, 29 April). *Web services architecture requirements*. Retrieved November 2002 from http://www.w3.org/TR/2002/WD-wsa-reqs-20020429

Banerji, A., Bartolini, C., Beringer, D., Chopella, V., Govindarajan, K., Karp, A., et al. (2002, March 14). *WSCL Web services conversation language*. Retrieved November 2002 from http://www.w3.org/TR/wscl10

Bartel, M., Boyer, J., Fox, B., LaMacchia, B., & Simon, E. (2002, February 12). *XML signature syntax and processing*. Retrieved November 2002 from http://www.w3.org/TR/2002/REC-xmldsig-core-20020212

BEA Systems, Intalio, SAP AG, and Sun Microsystems (2002). Web Service Choreography Interface (WSCI) 1.0 Specification. Retrieved November 2002 from http://wwws.sun.com/software/xml/developers/wsci

Berners-Lee, T. (1996, August). *The World Wide Web: Past, present and future*. Retrieved November 2002 from http://www.w3.org/People/Berners-Lee/1996/ppf.html

Chaum, D. (1985). Security without identification: Transaction systems to make Big Brother obsolete. *Communications of the ACM*, 28.

CryptoLogic Ecash FAQ (2002). Retrieved November 2002 from http://www.cryptologic.com/faq/faq-ecash.html

DAML: The DARPA Agent Markup Language Homepage (2001). Retrieved November 2002 from http://www.daml.org

Digital Certificates, CCITT (1988). Recommendation X.509: The Directory—Authentication Framework.

ebXML: Enabling a global electronic market (2001). Retrieved November 2002 from http://www.ebxml.org

Ford, W., Hallam-Baker, P., Fox, B., Dillaway, B., LaMacchia, B., Epstein, J., & Lapp, J. (2001, March 30). *XML key management specification (XKMS)*. Retrieved November 2002 from http://www.w3.org/TR/xkms

Glassman S., Manasse, M., Abadi, M., Gauthier P., Sobalvaro, P. (2000). The Millicent Protocol for Inexpensive Electronic Commerce. Retrieved November 2002 from http://www.w3.org/Conferences/WWW4/Papers/246/

Graupner, S., Kotov, V., & Trinks, H. (2002). Resource-sharing and service deployment in virtual data centers. In *IEEE Workshop on Resource Sharing in Massively Distributed Systems (RESH'02)*. New York: IEEE.

Hallam-Baker, P., & Maler, E. (Eds.). (2002, March 29). *Assertions and protocol for the OASIS Security Assertion Markup Language*. Retrieved November 2002 from http://www.oasis-open.org/committees/security/docs/draft-sstc-core-29.pdf

Karp, A., Gupta, R., Rozas, G., Banerji, A. (2001). The Client Utility Architecture: The Precursor to E-speak, HP Technical Report. Retrieved November 2002 from http://lib.hpl.hp.com/techpubs/2001/HPL-2001-136.html

HP Utility Data Center: Enabling the adaptive infrastructure. (2002, November). Retrieved November 2002 from http://www.hp.com/go/hpudc

Kim, W., Graupner, S., & Sahai, A. (2002, January 7–10). A secure platform for peer-to-peer computing in the Internet. Paper presented at 35th Hawaii International Conference on System Science (HICSS-35), Hawaii.

Kuno, H., & Sahai, A. (2002). *My agent wants to talk to your service: Personalizing Web services through agents*. Retrieved November 2002 from http://www.hpl.hp.com/techreports/2002/HPL-2002-114

IBM Autonomic Computing (n.d.). Retrieved from http://www.research.ibm.com/autonomic/

Leymann, F. (Ed.) (2001). *WSFL Web services flow language (WSFL 1.0)*. Retrieved July 2003 from http://www.ibm.com/software/solutions/webservices/pdf/WSFL.pdf

Liberty Alliance Project (2002). Retrieved November 2002 from http://www.projectliberty.org/

Machiraju, V., Sahai, A., & van Moorsel, A. (2002). Web service management network: An overlay network for federated service management. Retrieved November 2002 from http://www.hpl.hp.com/techreports/2002/HPL-2002-234.html

Micropayments overview (2002). Retrieved November 2002 from http://www.w3.org/ECommerce/Micropayments/

Microsoft .NET Passport (2002). Retrieved November 2002 from http://www.passport.com/

OASIS PKI Member Section (2002). Retrieved November 2002 from http://www.pkiforum.org/

Organization for the Advancement of Structured Information Standards (OASIS) (2002). Retrieved November 2002 from http://www.oasis-open.org

Reagle, J. (Ed.) (2000, October 6). *XML encryption requirements*. Retrieved November 2002 from http://lists.w3.org/Archives/Public/xml-encryption/2000Oct/att-0003/01-06-xml-encryption-req.html

RosettaNet (2002). Retrieved November 2002 from http://www.rosettanet.org

Sahai, A., Machiraju, V., Sayal, M., Jin, L. J., & Casati, F. (2002). Automated SLA monitoring for Web services. Retrieved November 2002 from http://www.hpl.hp.com/techreports/2002/HPL-2002-191.html

Sahai, A., Machiraju, V., & Wurster, K. (2001, July). Monitoring and controlling Internet based services. In *Second IEEE Workshop on Internet Applications (WIAPP'01)*. New York: IEEE. [Also as HP Tech. Rep. HPL-2000—120.]

Semantic Web (2001). Retrieved November 2002 from http://www.w3.org/2001/sw/

SET Secure Electronic Transactions LLC (2002). Retrieved November 2002 from http://www.setco.org/

Simple Public Key Infrastructure (SPKI). (1999). *SPKI Certificate Theory* (RFC 2693).

Thatte, S. (2001). XLANG Web Services for Business Process Design. Retrieved November 2002 from http://www.gotdotnet.com/team/xml_wsspecs/xlang-c/default.htm

TSpaces: Intelligent Connectionware (1999). Retrieved November 2002 from http://www.almaden.ibm.com/cs/TSpaces/

Van Moorsel, A. (2001). *Metrics for the Internet Age—Quality of experience and quality of business*. Retrieved November 2002 from http://www.hpl.hp.com/techreports/2001/HPL-2001-179.html

Weber, R. (1998). *Chablis—Market analysis of digital payment systems*. Retrieved November 2002 from University of Munich Web site: http://chablis.informatik.tu-muenchen.de/MStudy/x-a-marketpay.html

Web Site Design

Robert E. Irie, *SPAWAR Systems Center San Diego*

INTRODUCTION

Designing and implementing a Web site is increasingly becoming a complex task that requires knowledge not only of software programming principles but of graphical and user interface design techniques as well. While good design is important in regular software engineering and application development, nowhere is it more essential than in Web site development, due to the diverse and dynamic nature of Web content and the larger intended audience.

This chapter will cover some of the issues involved with the two major components of a Web site, its design and implementation. The scope of this chapter is necessarily limited, as Web development is a rich and heterogeneous field. A broad overview of techniques and technology is given, with references to other chapters. The reader is directed to consult other chapters in this encyclopedia for more detailed information about the relevant technologies and concepts mentioned below. Occasionally links to Web sites will be given. They are either representative examples or suggestions for further reference, and should not be construed as an endorsement.

WEB SITE COMPONENTS

A Web site is an integration of three components, the content to be published on the Web, its presentation to the user, and the underlying programming logic. Each component has its own particular representation and role in shaping the overall user experience.

Content

The content consists of all relevant data that are to be published, or shown to the user. It usually constitutes the bulk of a Web site's storage requirements and can be in the form of text, images, binary and multimedia data, etc. Static textual and graphic content can be stored as HTML pages, whereas multimedia files like videos and sound recordings are usually stored in large databases and served, in whole or in parts, by dedicated servers. Most of the discussion in this chapter will focus on the former type.

Presentation

The presentation component involves the user interface to the Web site and the manner in which content is displayed. Typical elements include the graphical and structural layout of a Web document or page, text and graphic styles to highlight particular content portions, and a mechanism for the user to navigate the Web site. Originally, files with HTML markups were used to store both content and information regarding its presentation. It is now common practice to store neither exclusively in HTML. HTML is primarily used to describe the structure of a Web document, by breaking down the page into distinct elements like paragraphs, headings, tables, etc. The actual textual content of the document can be stored separately in a database, to be dynamically inserted into the HTML page using programming logic. A separate file, called the style sheet, can be associated with the HTML page, and consequently the content, to affect the presentation. A style sheet file can describe how each structural element in an HTML file is displayed; sizes, colors, positions of fonts, blocks, backgrounds, etc., are all specified in a hierarchical organization, using the standard Web-based style sheet language, cascading style sheets (CSS) (Lie & Bos, 1999).

Logic

The programming logic determines which content to display, processes information entered by the user, and

generates new data. It drives the interaction between the Web site and the user and is the glue that binds the content and its presentation. To be useful, it needs to access the content as well as its presentation information and handle user input accordingly. Logic is usually implemented as small programs, or scripts, that are executed on the Web server or the user's browser. These scripts can be stored within the HTML page, along with the presentation and content, or separately as distinct program files that are associated with the content. There are several standard programming languages that can be used in writing scripts.

Separation of Components

With the existence of a variety of technologies, protocols, and standards, Web development is remarkably flexible, and there are often multiple ways of accomplishing the same task. This is both an asset and a liability, as while developers are free to choose their own techniques, it is very easy to create sloppy or undisciplined documents and code. In regular application development, it is important to adhere to sound software engineering techniques to manage a code base for future enhancements and simultaneous development efforts. The flexibility of Web development makes such good techniques even more critical.

Until very recently, there was a great deal of overlap, in terms of storage and implementation, of the three Web site components mentioned above. This led, for example, to Web pages that contained all three components in a single, often unmanageable, HTML file. As Web site development has matured, the principle of Web site component separation has become widely encouraged, if not accepted, and it is the central theme of this chapter.

IMPLEMENTATION ISSUES

The World Wide Web (WWW) is a series of client/server interactions, where the client is the user's Web browser, and the server is a particular Web site. The WWW Consortium (W3C) defines the hypertext markup language (HTML) and the hypertext transfer protocol (HTTP) as the standard mechanisms by which content is published and delivered on the Web, respectively.

In essence, the local Web browser initiates HTTP requests to the remote Web server, based on user input. The Web server retrieves the particular content specified by the requests and transmits it back to the browser as an HTTP response. The Web browser then interprets the response and renders the received HTML content into a user-viewable Web page.

Web site implementations can be classified by the level of interactivity and the way content is stored, retrieved, and displayed.

Static Sites

Static sites are the simplest type of Web sites, with the content statically stored in HTML files, which are simple text files. Updating the Web site requires manually changing individual HTML text files. While this type of site was prevalent in the beginning, most sites, especially commercial ones, have come to incorporate at least some degree of dynamic behavior, and users have come to expect some interactivity.

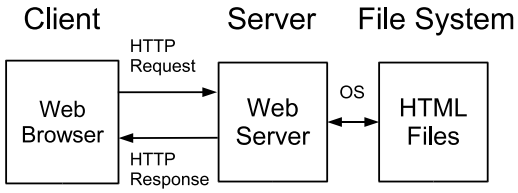


Figure 1: Block diagram of a client/server architecture with a static Web site.

Figure 1 shows the basic client–server interaction for a static Web site. The client browser makes an HTTP request to a Web server. The URL specifies the particular Web server and page. The Web server retrieves the requested Web page, which is an HTML file, from the file system and sends it back to the client through the HTTP response.

This very basic interaction between browser and server is the basis for more complex, dynamic interactions. This type is static because the Web page content is straight HTML, statically stored on disk, with no mechanism to change the contents. The Web server here serves solely as a file transfer program.

Developing static sites requires very few tools. All that is required, besides the Web server and browser, is a text editor application. The simplest text editor can be used to manually create HTML files. Complex, graphical HTML editors can make the task almost trivial by automatically generating HTML files and behaving similarly to word processors, with WYSIWYG (what you see is what you get) interfaces. Creating graphics and images for static Web sites is also straightforward, requiring any typical paint or drawing program.

DYNAMIC SITES

Dynamic sites share the same basic architecture as static ones, but with the addition of programming logic. The two major types of dynamic sites reflect the place of execution of the scripts. Client-side scripting involves embedding actual source code in HTML, which is executed by the client browser in the context of the user's computer. Server-side scripts, on the other hand, are executed on the Web server. While the following discussion examines both types separately, in an actual Web site both types can and often do exist simultaneously.

Client Side

Figure 2 shows the basic architecture for a dynamic site with client-side scripting. Scripts are embedded within HTML documents with the <script> </script> tags or stored in separate documents on the server's file system. Scripts are transmitted, without execution, to the client browser along with the rest of the HTML page. When the client browser renders the HTML page, it also interprets and executes the client script. An example of a client-side script is the functionality that causes a user interface element, such as a menu, to provide visual feedback when the user moves the mouse pointer over a menu option.

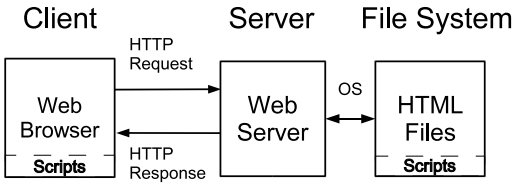


Figure 2: Block diagram of a Web site interaction with client-side scripting.

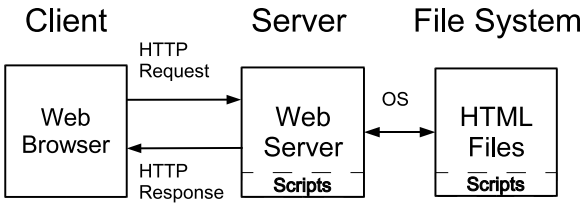


Figure 3: Block diagram of a Web site interaction with server-side scripting.

There are several client-side scripting languages, the most common one being JavaScript, an object-oriented language originally developed by Netscape. It is now a standardized language, defined by the international industry group European Computer Manufacturers Association, and called ECMAScript (European Computer Manufacturers Association, 1999). Netscape continues to use the term JavaScript, however, and Microsoft calls its implementation of ECMAScript for Windows browsers JScript. The other major scripting language is Microsoft's VBScript, short for Visual Basic Scripting Edition, which is available only for Windows platforms (Champeon, 2001).

Regardless of the language, client-side scripts rely on a standard programming interface, defined by the W3C and called the Document Object Model (DOM), to dynamically access and update the content, structure, and style of Web documents (World Wide Web Consortium, 1998).

Cascading style sheets (CSS) is another W3C language standard that allows styles (e.g., fonts, colors, and spacing) to be associated with HTML documents. Any specific HTML tag or group of HTML tags can be modified. It is a language, separate from HTML, that expresses style in common desktop publishing terminology. The combination of HTML, CSS, and DOM client-side scripts is often referred to as dynamic HTML (Lie & Bos, 1999).

Client-side scripting is used primarily for dynamic user interface elements, such as pull-down menus and animated buttons. The advantage of using client-side scripts instead of server-side scripts for such elements is that the execution is more immediate. Since the script, once loaded from the server, is being executed by the browser directly on the user's computer, there are no delays associated with the network or the server load. This makes the user interface responsive and similar to standard platform applications.

One of the disadvantages is that client-side scripting languages are usually more limited in functionality than server-side languages, so that complex processing is not possible. Such limitations are by design, for security reasons, and are not usually apparent for simple user interface programming.

Users may also specifically choose not to allow client-side scripts to execute on their computers, resulting in a partial or complete reduction in functionality and usability of a Web site. In general, it is recommended that a site incorporate user interface scripting only sparingly, and always with clear and functional alternatives.

Finally, because client-side programs, whether embedded or stored separately, must necessarily be accessible and viewable by the Web browser, they are also ultimately viewable by the user. This may not be desirable for commercial Web applications, where the programming logic can be considered intellectual property.

Since client-side scripts are embedded in HTML pages, any tool that creates and edits HTML pages can also be used to create the scripts. The only requirement is that the client browser support the particular scripting language. Most modern browsers support some variation of Javascript/ECMAScript/JScript, whereas a smaller subset support VBScript.

Server Side

Figure 3 shows the basic architecture for a server-side dynamic site. Scripts are still stored in HTML documents on the server's file system, but are now executed on the server, with only the program results and output being sent to the client browser, along with the rest of the HTML page. To the client browser, the HTTP response is a normal static HTML Web page. Scripts are embedded in HTML documents using special HTML-like tags, or templates, whose syntax depends on the particular server-side scripting language (Weissinger, 2000).

There are several common server-side scripting languages, including PHP, Active Server Pages (ASP), and Java Server Pages (JSP). The common gateway interface (CGI) is also a server-side scripting mechanism, whereby neither the Web content nor the programming logic is stored in an HTML file. A separate program, stored in the file system, dynamically generates the content. The Web server forwards HTTP request information from the client browser to the program using the CGI interface. The program processes any relevant user input, generates an HTML Web document and returns the dynamic content to the browser via the Web server and the CGI interface. This process is illustrated in Figure 4.

Server-side scripting is used primarily for complex and time-consuming programming logic tasks, where immediacy of response is not as critical as with user interface elements. The advantage of using server-side scripts is the freedom and computational power that is available on the

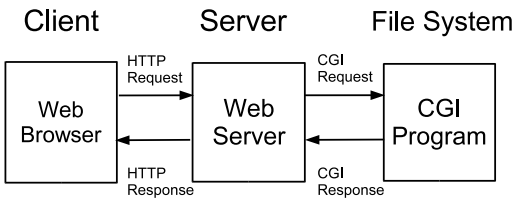


Figure 4: Block diagram of a Web site interaction with common gateway interface scripting.

server; server-side scripts do not have the same security constraints as client-side scripts, and often have full access to the server machine's file system and resources. The user may not disable execution of such scripts, so that the Web developer can reasonably expect that the Web site will behave exactly the same regardless of user configuration. Finally, any proprietary server-side source code is safely hidden from user view, as the client browser receives only the output of the script.

Server-side scripts have the disadvantage of requiring a request–response round trip between the client browser and the server, which leads to slower response times.

Server-side scripting languages normally interact closely with the Web server, which imposes some compatibility constraints. The choice of a Web server, particularly a proprietary system, usually limits the selection of server-side scripting languages, and vice versa.

WEB APPLICATIONS

As a Web site becomes more complex, a robust and efficient mechanism for the separation of content, presentation, and logic is necessary. Web application servers are Web sites that are more interactive, access large amounts of data, and provide a rich functionality similar to that of desktop applications. Unlike desktop applications, where all components are stored and executed on the same computer, Web applications usually follow a three-tier client/-server architecture (see Figure 5) consisting of the Web browser, the Web server, and a database. All content and logic are stored in the database and are retrieved and processed as necessary on the Web server. The presentation information can be embedded with the content or stored as a separate style sheet on the database or the server.

Usually a Web application server interfaces with a relational database, which stores data in rows of tables.

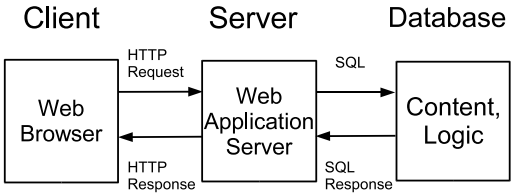


Figure 5: Block diagram of a Web application server interaction.

The other major type of database is the object-oriented database, which stores data by encapsulating them into objects. Relational databases are often more efficient and faster than equivalent object-oriented databases and support an almost universal database language, SQL (structured query language).

The major disadvantage of developing with Web application servers, besides the inherent complexity, is the necessity of learning a nonstandard or proprietary server-side programming interface or language. There are several major Web application servers that support standard programming languages such as Java and C++, but each has its own application programming interface (API). Table 1 lists some of the popular commercial and open source application servers (see Web Resources).

DESIGN ISSUES

Unlike implementation issues, which usually are straightforward to specify and quantify, design issues are much more subjective and are dependent on several factors, including the particular type of Web site and its purpose. Web site development efforts are often driven by conflicting objectives and considerations, and a balance must be maintained between business and financial concerns,

Table 1 URLs of Various Web Resources

Browsers	Link
Internet Explorer	http://www.microsoft.com/windows/ie
Netscape Navigator	http://browsers.netscape.com/browsers
Lynx	http://lynx.browser.org
Design Guidelines	
Fixing Your Web site	http://www.fixingyourwebsite.com
CSS	http://www.glish.com/css
Usability and Accessibility Issues	http://usability.gov/index.html
Programming	
DevShed	http://www.devshed.com
Webmonkey	http://www.webmonkey.com
Javascript	http://www.javascript.com
Standards	
World Wide Web Consortium	http://www.w3.org
Web Application Servers	
BEA WebLogic	http://www.beasys.com/products/weblogic
IBM WebSphere	http://www.ibm.com/software/webservers/appserv
Macromedia ColdFusion	http://www.macromedia.com/software/coldfusion
Apache Jakarta	http://jakarta.apache.org
Zope	http://www.zope.org

which often stress the commercial viability and revenue-generating aspects of a Web site, and more user-centric design concerns, which usually deal with usability issues (Murray & Costanzo, 1999). Since the former are very domain-specific, only the latter will be discussed in this chapter. In the discussion that follows, references to sample Web sites will be given.

USABILITY ISSUES

The goal of designing a Web site with usability issues in mind is to ensure that the users of the site find it usable and useful. Specifically, a Web site should be accessible, appealing, consistent, clear, simple, navigable, and forgiving of user errors (Murray & Costanzo, 1999).

The first step in designing any Web site should be the determination of the purpose of the site. Too often the rush to incorporate the latest Web technology or standard prevents a thorough examination and determination of the most important factor of the Web site, its intention or purpose. Most Web sites in essence are information dissemination mechanisms; their purpose is to publish useful content to as wide an audience as possible. Others also have a commercial component, with the buying and selling of goods or services. Still others foster a community or group activity and are used as collaboration devices.

The Web site's purpose should drive the design and implementation efforts. A Web site advertising or describing a company's products will most likely need eye-catching graphical designs and images. A commerce site will need to consider inventory mechanisms and secure transactions. A community site will need to solve problems involving simultaneous collaboration of a distributed group of users.

It is also important to consider the intended audience of a Web site. There is a wide range in browser capabilities and user technical competencies that must be taken into account. A Web site geared toward a younger, more technically inclined audience may contain highly interactive and colorful designs, whereas a corporate site might want to have a more professional, businesslike appearance. It is generally a good practice, if not essential, to consider accessibility issues for all users, including those who do not have access to high-end graphics-capable browsers.

BASIC WEB SITE TYPES

Just as there are several implementation classifications for Web sites, we can also classify them based on their purpose. Each type will lead to different choices in the content, presentation, and logic components and require emphasis on different usability issues. A single Web site may incorporate features of more than one basic type.

News/Information Dissemination

This type of Web site design is geared toward providing informational content to the Web user. The content is usually textual in form, with some graphics or images. The presentation of the content and its navigation are kept as clear and consistent as possible, so that the user will be able to quickly access the desired information. Not surprisingly, newspaper companies usually have Web sites with online news content (e.g., http:/./www.nytimes.com).

Portal

A portal is a popular type of Web site that serves as a gateway to other Web sites. The content is usually in the form of URL links and short descriptions, categorized based on themes. The links should be organized so that they are easily searchable and navigable. Major commercial portals have evolved from simple collections of related URL links to incorporate more community-like features to prompt users to return to their sites (e.g., http://www.yahoo.com).

Community

Community sites foster interaction among their users and provide basic collaboration or discussion capabilities. Message boards, online chats, and file sharing are all typical functionalities of community sites. The open source software movement has promoted numerous Web sites based on this type (e.g., http://www.sourceforge.net).

Search

There is a lot of overlap between this type of Web sites and portals. Like portals, search sites provide a mechanism by which users discover other Web sites to explore. Some sophisticated programming logic, the search engine, forms the foundation of this type of Web site. Search sites often emphasize simple, almost minimalist interfaces (e.g., http://www.google.com).

E-commerce

This type of site is often a component of other Web site types and allows users to purchase or sell goods and services in a secure manner. Since potentially large amounts of currency are involved, security is an important consideration, as well as an interface that is tolerant of potential user errors. An example of a successful commerce site with elements of a community is http://www.ebay.com.

Company/Product Information

With widespread Web use, having an official Web presence is almost a requirement for corporations. Such sites usually serve purposes similar to those of informational and e-commerce sites, but with a more focused interface, reflecting the corporate image or logo (e.g., http://www.microsoft.com).

Entertainment

This type of site is usually highly interactive and stresses appealing, eye-catching interfaces and designs. Typical applications include online gaming sites, where users may play games with each other through the site, and sporting event sites, where users may view streaming content in the form of video or audio broadcasts of live events (e.g., http://play.games.com).

BASIC DESIGN ELEMENTS

There is obviously no single best design for a Web site, even if one works within a single type. There are, however, some guidelines that have gained common acceptance. Like any creative process, Web site design is a matter of tradeoffs. A typical usability tradeoff is between making

an interface appealing and interactive and making it clear and simple. The former usually involves graphical designs with animations and client-side scripting, whereas the latter favors minimal text-based interfaces. Where a particular Web site belongs on the continuous spectrum between the two extremes depends on its intended purpose and audience, and should be a subjective, yet conscious decision.

The safest design decision is to always offer alternatives, usually divided into high- and low-bandwidth versions of the same Web pages, so that the user experience can be tailored to suit different preferences. The major disadvantage of this is the increase in development time and management requirements.

Accessibility/Connectivity

The two major factors affecting accessibility and connectivity issues are the bandwidth of the user's network connection, and the particular graphical capabilities of the user browser. Low-bandwidth connections to the Internet are still very common in homes. By some measures, dialup modems are still used in 90% of all homes that regularly access the Internet (Marcus, 2001). This requires Web site designers either to incorporate only small, compressed images on their sites, or to provide alternative versions of pages, for both high- and low-bandwidth users.

Some user browsers do not have any graphics capability at all, for accessibility reasons or user preference. For example, visually impaired users and PDA (personal digital assistant) users most often require accessibility consideration. Estimates of the number of disabled users range from 4 to 17% of the total online population (Solomon, 2000). PDA and mobile phone Internet usage is relatively new in the United States, but is already approaching 10 million users (comScore Networks, 2002). For such users, designing a separate text-only version of the Web site is a possibility. What would be better is to design a Web site that contains automatic browser-specific functionality degradation. An example is to associate relevant textual content to graphical images; graphical browsers may display the images, while text browsers may display the descriptions.

Consistent Page Layout

One of the most important design for a Web site is a consistent page layout. While every single page does not need to have the same layout, the more consistent each page looks, the more straightforward it is for the user to navigate through the site and the more distinctive the Web site appears. A typical Web page layout utilizes parts or all of an artificially defined border around the content (see Figure 6).

Originally, HTML frames or tables were the standard way of laying out a page, and they are still the preferred method for most developers. However, the W3C clearly is favoring the use of cascading style sheets (CSS) for page layout (World Wide Web Consortium, 2002). CSS also provides a mechanism for associating styles, such as color, font type and size, and positioning, with Web content, without actually embedding them in it. This is in keeping with the principle of separating content from its presentation.

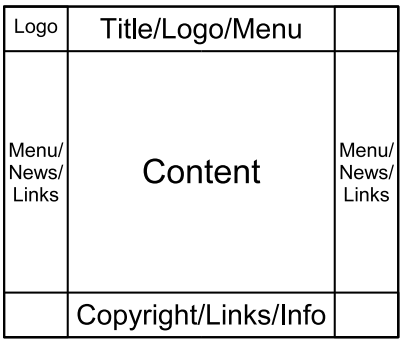


Figure 6: A typical layout scheme for a Web page.

Consistent Navigation Mechanism

Web site navigation is an important component of the design, and a consistent navigation mechanism supplements a page layout and makes the user experience much simpler and more enjoyable.

One of the best ways of providing a consistent navigation mechanism is to have a menu bar or panel that is consistent across all pages of the site. Such a menu can be a static collection of links, or a dynamic, interactive component similar to that of a desktop application. Figure 7 is an example of a simple and consistent navigation scheme that utilizes two menu panels. The top panel (with menu items A, B, C) is similar to a desktop application's menu bar and is a global menu that is consistent throughout the site and refers to all top-level pages of a site. The left side panel is a context-dependent menu that provides further options for each top-level page. This type of navigation scheme can be seen on several public Web sites (e.g., http://www.python.org).

While there are no absolute rules or guidelines for good navigation elements, they usually provide visual feedback (e.g., mouse rollover effects), have alternate text displays (for nongraphical or reduced capability browsers), and are designed to be easy to use as well as learn.

MISCELLANEOUS INTERFACE ISSUES

The following are miscellaneous interface design issues. Again, only suggestions for possible design choices are offered, and they will not be applicable in all circumstances.

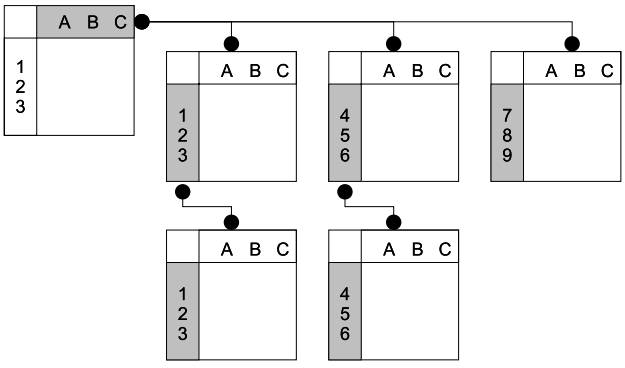


Figure 7: An example of a consistent navigation scheme for a Web site.

Graphics

The two major interface issues concerning graphics are size and color. As the majority of Web users still access Web sites over slow modem links, it is important to use graphic images that are of reasonable size, to prevent excessive download delays. For photograph images, the JPEG format offers a good compromise between lossy compression size and image quality, with an adjustable tradeoff point. For line art and solid color images, lossless compression is preferred, and the proprietary GIF format is common, although the open standard PNG format is gaining in acceptance (Roelofs, 2000).

The issue of colors is a complex one and depends on many factors. In general, Web graphic designers work with 24-bit colors, with 256 possible values for each of three color channels, red, green, and blue (RGB). Until recently, the majority of users' computers and Web browsers could only support a palette, or set, of 256 colors simultaneously. To ensure that colors appear uniformly across platforms and browsers, a "Web-safe palette" of 216 colors was established, consisting of combinations of six possible values, or points, for each of three color channels (6 possible reds $\times$ 6 possible greens $\times$ 6 possible blues = 216 possible colors) (Niederst, 2001).

Recently, browsers and systems with 24-bit and 16-bit support have drastically increased and now account for about 94% of all users (Lehn & Stern, 2000). Twenty-four-bit color support results in the full display of the designer's color scheme. Sixteen-bit color systems are sometimes problematic, as they nonuniformly sample the three color channels (5 bits for red, 6 bits for green, and 5 bits for blue) and provide a nonpalettized approximation of 24-bit color.

Layout Styles

A comprehensive guide to layout styles is beyond the scope of this chapter. The major design decision is between having page layouts of fixed or variable size (Niederst, 2001). By default, HTML documents are variable-sized, in that text and graphics positioning and line breaks are not determined by the user's monitor resolution and browser window size. Since a wide variety of sizes and resolutions is almost a given, having a variable-sized page layout allows flexible designs that scale to the capabilities and preferences of each user. The disadvantage is that because each user experience is different, and elements can be resized or repositioned at will, it is difficult to design a consistent and coherent interface; there is the possibility that some configurations lead to poor or unusable interfaces.

The alternative to the default variable-sized page layout is to explicitly design the size and position of some or all of the elements of a Web document. An example of this would be to limit the width of all content in a page to fit within a certain horizontal screen resolution, such as 640 pixels. All text and graphics will remain stationary even if the user resizes the browser window to greater than 640 horizontal pixels. The advantage of this method is that designing an interface is much more deterministic, so the Web designer will have some degree of control over the overall presentation and is reasonably certain that all users will have the same experience accessing the site. The disadvantage is that the designer must pick constants that may not be pleasant or valid for all users. For example, a Web page designed for a 640 $\times$ 480 resolution screen will look small and limited on a 1280 $\times$ 1024 screen, whereas a Web page designed for an 800 $\times$ 600 screen would be clipped or unusable for users with only a 640 $\times$ 480 screen.

Actually implementing either type of page layout can be done with HTML frames, tables, or CSS style sheets, or some combination of the three. Although using style sheets is the currently preferred method for page layout, browser support is still poor, and many sites still use frames or tables (Niederst, 2001).

Search Engines

A search engine is a useful tool to help users quickly find particular content or page as the content of a Web site increases, or the navigation scheme becomes complicated. The search engine is a server-side software program, often integrated with the Web server, that indexes a site's Web content for efficient and quick retrieval based on a keyword or phrase. Search engines are available with a variety of configurations, interfaces, and capabilities. A good resource that summarizes the major commercial and open source engines is the Search Tools Web site (http://www.searchtools.com).

Cross-Browser Support

Designing a Web site that is consistent across multiple browsers and platforms is one of the most challenging tasks a developer faces. Even different versions of the same browser are sometimes incompatible. At the minimum, the three major browsers to consider are Internet Explorer (IE), Netscape Navigator (NN), and text-based browsers such as Lynx.

For the most part, browser development and capabilities have preceded the establishment of formal standards by the W3C, leading to numerous incompatibilities and nonuniform feature support. The latest versions of the two common browsers (IE 6, NN 6.2) offer complete support for the current W3C standard HTML 4.01. However, the more common, earlier versions of the browsers (versions 4+ and 5+) had only incomplete support.

Even more troublesome was their support of the W3C standard Document Object Model (DOM) Level 1, as each has historically taken a different track and implemented its own incompatible DOM features (Ginsburg, 1999). In general, NN's DOM support is much closer to the "official" W3C DOM Level 1 specification, whereas IE has several extensions that are more powerful, but are available only on Windows platforms. The latest versions of the two browsers have alleviated some of this problem by supporting, as a baseline, the complete Level 1 specification.

WEB RESOURCES

Table 1 summarizes some useful online resources for Web site development. They are only suggestions and should not be considered comprehensive or authoritative.

CONCLUSION

This chapter has given an overview of Web site development, including the design and implementation aspects.

This field is very dynamic, and technologies and practices are constantly changing. More complex object-oriented programming paradigms and generalized markup languages are gaining widespread acceptance and use. XML (extensible markup language), XHTML (extensible HTML), XML-RPC (XML remote procedure call), SOAP (simple object access protocol), and SVG (scalable vector graphics) are examples of such new standards. However, the basic principles of clarity, consistency, and conciseness are still applicable to the design of all sites regardless of type or technology.

The Web development profession is also rapidly changing field. No longer is it feasible to have one person perform all design and implementation duties. A team of Web developers, graphic designers, and database administrators is usually required, with each member responsible for the three components of Web site development: content management, content presentation, and programming logic. However, it is still important to be aware of all issues in order to work effectively in a Web development team.

GLOSSARY

Client/server architecture A process by which multiple computers communicate. The client initiates all communication with the server in the form of a request and receives the results in the form of a response. For Web sites, the user's browser is the client requesting content or services from the Web server.

Database A repository of information. The data are stored in a structured way to be easily and efficiently retrieved. Two popular types of databases are the relational database and the object-oriented database. Each has advantages and disadvantages with respect to efficiency, rich associations between information, etc.

Hypertext A mechanism by which related content (text, graphic, multimedia, etc.) is associated using links. A hypertext document allows the user to easily access relevant content in a seamless, integrated context, as opposed to traditional, sequentially viewed documents.

Hypertext markup language (HTML) A standard language for publishing content on the World Wide Web. HTML defines a set of markups, or tags, that are embedded in a Web document and provide structural, stylistic, and content information.

Uniform resource locator (URL) The explicit format for a reference to a hypertext document. It is in the form *protocol://server:port/path*. The protocol can be any of several standard Internet communications protocols, with HTTP being the most common for Web pages. By default, Web servers communicate using a standard port number, 80. In such cases the URL can be shortened to *protocol://server/path*.

User Anyone accessing the Web site, using a Web browser. A related term, user interface, refers to the entire environment (text, graphics, and user input and response) that builds the experience for the user interacting with the site.

Web site The integration of hypertext content, presentation information, and controlling logic, that forms the user experience. Implemented on a Web server, its purpose is usually to disseminate information, foster collaboration, or obtain user input. It is the basic unit of discussion in this chapter and will refer to both the user experience and the actual implementation.

World Wide Web (WWW) A network of hypertext documents, existing on Web servers and accessible via the Internet using computer programs called Web browsers.

CROSS REFERENCES

See *Client/Server Computing; Databases on the Web; HTML/XHTML (HyperText Markup Language/Extensible HyperText Markup Language); Usability Testing: An Evaluation Process for Internet Communications*.

REFERENCES

Champeon, S. (2001). JavaScript: How did we get here? Retrieved April 16, 2002, from http://www.oreillynet.com/pub/a/javascript/2001/04/06/js_history.html

comScore Networks (2002). Ten million Internet users go online via a cellphone or PDA, reports Comscore Media Metrix Press Release. Retrieved August 30, 2002, from http://www.comscore.com/news/cell_pda_082802.htm

European Computer Manufacturers Association (1999). Standard ECMA-262: ECMAScript language specification. Retrieved April 2, 2002, from ftp://ftp.ecma.ch/ecma-st/Ecma-262.pdf

Ginsburg, P. E. (1999). Building for 5.0 browsers. Retrieved May 10, 2002, from http://builder.cnet.com/webbuilding/pages/Authoring/Browsers50

Lehn, D., & Stern, H. (2000). Death of the websafe color palette? Retrieved April 20, 2002, from http://hotwired.lycos.com/webmonkey/00/37/index2a.html?tw=design

Lie, H. W., & Bos, B. (1999). Cascading style sheets, level 1 W3C recommendation. Retrieved April 2, 2002, from http://www.w3.org/TR/REC-CSS1

Marcus, B. (2001). Wireless, broadband penetration continues. Retrieved May 1, 2002, from http://www.digitrends.net/nwna/index_15935.html

Murray, G., & Costanzo, T. (1999). Usability and the Web: An overview. Retrieved April 16, 2002, from http://www.nlc-bnc.ca/9/1/p1-260-e.html

Niederst, J. (2001). *Web design in a nutshell* (2nd ed.). Sebastopol, CA: O'Reilly & Associates.

Roelofs, G. (2000). PNG, MNG, JNG, and Mozilla M17. Retrieved August 30, 2002, from http://www.libpng.org/pub/png/slashpng-2000.html

Solomon, K. (2000). Smart biz: Enabling the disabled. Retrieved May 1, 2002, from http://www.wired.com/news/print/0,1294,39563,00.html

Weissinger, A. K. (2000). *ASP in a nutshell* (2nd ed.). Sebastopol, CA: O'Reilly & Associates.

World Wide Web Consortium (1998). Document object model (DOM) level 1 specification. Retrieved April 2, 2002, from http://www.w3.org/TR/REC-DOM-Level-1

World Wide Web Consortium (2002). Hypertext markup language (HTML) home page. Retrieved August 28, 2002, from http://www.w3.org/MarkUp

Wide Area and Metropolitan Area Networks

Lynn A. DeNoia, *Rensselaer Polytechnic Institute*

INTRODUCTION

In today's social, political, and economic environment, individuals and organizations communicate and operate over ever-increasing geographic distances. This means that access to and sharing of information and resources must extend beyond the "local" office, building, or campus out across cities, states, regions, nations, continents, and even beyond the planet. Bridging this diversity of distances in ways that satisfy application requirements for speed, capacity, quality, timeliness, etc. at reasonable cost is no simple challenge, from either a technical or a business perspective. In this chapter we concentrate on the main elements required to meet such a challenge in wide area and metropolitan area networks.

HISTORY AND CONTEXT

Definitions

The public networking arena has typically been divided into two segments with the following characteristics:

Metropolitan area networks (MANs) are built and operated by service providers (SPs) who offer network services to subscribers for a fee, covering distances up to tens of miles, often within or surrounding a major city. MANs are often built by telecommunication companies such as local exchange carriers (LECs) or by utility companies. A recent alternative using Ethernet for the MAN has spawned a new category of companies called Ethernet LECs or ELECs.

Wide area networks (WANs) are built and operated by SPs who offer network services to subscribers for a fee, covering distances up to hundreds or thousands of miles, such as between cities, across or between countries, across oceans, etc. WANs designed for voice are usually built by telecommunication companies known in the United States as interexchange carriers (IXCs). WANs for data are also called public data networks (PDNs).

By contrast, local area networks (LANs) are typically built and operated as private networks, by individuals or enterprises, for their own use. In addition, landlords operating as building LECs (BLECs) may offer LAN services to tenants. In either case, the geographic scope of a LAN is usually limited to a building or campus environment where all rights of way for cabling purposes belong to the individual/enterprise/landlord. The boundaries between LANs and MANs and WANs began to blur as geographic limitations of networking technologies were extended with increasingly capable implementations over fiber-optic cabling. Even the distinctions between private and public networks became more difficult to draw with the advent of "virtual private network" equipment and services.

Challenges

The number of options and choices available to network designers in both the subscriber and provider communities continues to grow for both MANs and WANs. Multiple technologies and standards, increasing numbers and types of applications, higher expectations for MAN and WAN performance comparable to (or at least approaching) that found in a LAN environment, and pressure to keep unit costs low, all combine to create enormous challenges for MAN and WAN builders. Infrastructure choices must last long enough, not just for cost recovery, but to achieve return on investment. Service providers must marry new technologies to their existing installed base, create smooth transitions (e.g., for network upgrades, new service roll-outs) with minimal disruption to customer services, and add or enhance services to meet advancing customer expectations, all in an environment of increasing economic and competitive pressure. Many providers have begun to recognize that their long-term survival depends on a strategy of simplification—reducing the complexity (to have fewer technologies, fewer equipment vendors, fewer equipment types, fewer management systems, etc.) of their infrastructure while maintaining the flexibility to adapt to changing application, user, and competitive requirements.

The pressure to simplify is constantly at odds with the difficulties of predicting the future:

Which technology will provide the best flexibility and scalability at an acceptable investment cost?

How fast and in what ways will application needs and user expectations develop?

Which services or enhancements will provide competitive advantage?

How can value be added to those elements moving downward into the commodity market?

The ability to develop shrewd answers to such questions is likely to determine which companies will thrive in the networking services business.

Functional Requirements

The basic function that subscribers seek from MAN and WAN service providers is the ability to deliver traffic from one place to another (point-to-point) or to multiple others (multipoint). This begins with *connectivity*. For the network in Figure 1, traffic can flow from A to C and/or D, but not to B. Once connectivity is established, the network must have sufficient *capacity* in bandwidth and switching to get the traffic from the source to its intended destination. Subscribers want services that are reliable, as measured by the percentage of time network resources are available when needed and by the amount of traffic (preferably none) that gets lost. Subscribers also want services that perform well enough so that their traffic gets delivered in a timely fashion, with minimal delay (low latency is particularly important for delay-sensitive traffic such as voice or video). Providers, on the other hand, want an infrastructure that is cost-effective, manageable, and capable of supporting revenue generation and profits.

Evolution and Coexistence

The first WANs were built from circuit-switched connections in the telephone system because that's what was available to cover the distances involved. Circuit switching continues to be useful, particularly when the computer

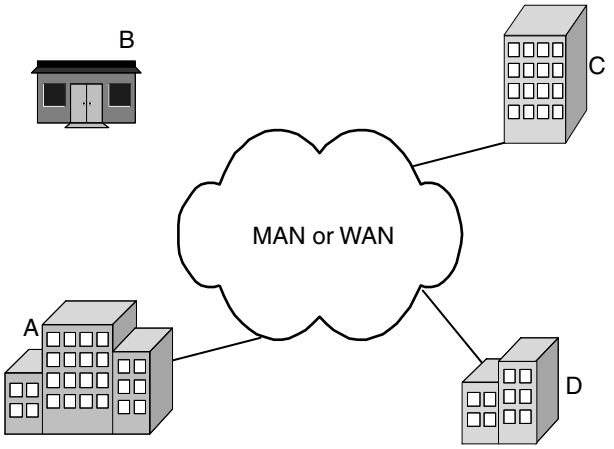


Figure 1: Connectivity.

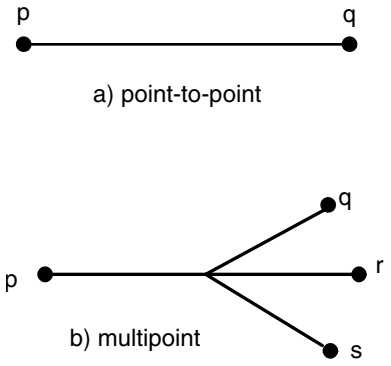


Figure 2: Connections, a) point-to-point and b) multipoint.

devices being connected need to exchange messages in real time or with guaranteed delivery. For occasional traffic, dial-up connections similar to an individual telephone call are used. For continuous traffic or when applications cannot tolerate the delay involved in call setup, circuits are leased from a telephone company and "nailed up" into permanent connections. For two connected locations the leased line is called a *point-to-point* connection (Figure 2a). More than two locations can be connected with a *multipoint* link (Figure 2b) if a sharing discipline is imposed to prevent traffic from one source interfering with traffic sent from another at the same time. In either case, the resources required to carry traffic across the leased line are dedicated to the particular subscriber, creating an effectively private connection through the service provider's public network resources.

Two devices connected by a leased line may or may not send traffic continuously, wasting capacity when the line is idle. If there are multiple devices in one location to be connected to one or more devices in a destination location, a single leased line may be shared using a device at each end of the line called a multiplexer. Statistical multiplexing allows more devices to be connected than the capacity of the line could support in real time if all were to transmit simultaneously. This is called oversubscription. On the average, it is quite likely that only some devices will be active, and the line is shared effectively with little traffic delay and less wasted capacity. However, when many devices are active, performance can be degraded. The sending multiplexer adds a label to each unit of traffic transmitted; the receiver reads (and removes) the label to figure out which device is the intended recipient and switches the traffic onto the appropriate output link. Packet switching is a form of statistical multiplexing.

Originally circuit switching was designed to carry analog voice traffic and packet switching was designed for digital data. Today, however, public networks convert all types of traffic into digital form for cost-effective transport. We could say that "bits are bits," whether they belong to voice, data, video, or some other application. The same network might well be used to deliver multiple types of bits, instead of having distinct networks dedicated for voice, data, etc. This is the concept of *convergence*, where a single network carries various types of traffic. In the context of convergence, the important question shifts from whether circuit or packet switching is better, to what support a network

must provide so that traffic delivery meets user expectations and application requirements. Convergence is certainly not new, because in early WANs, digital data were transformed into analog signals and carried over public networks that had been designed for voice. Today convergence is available through many more options for what traffic to combine and how to do it.

FACILITIES AND INFRASTRUCTURE

Digital Transmission

The heritage of digital WANs dates from the early 1960s, when the Bell System first introduced the T-carrier system of physical components to support transport of digital signals in the United States. The accompanying time-division multiplexed (TDM) digital signal scheme, called a digital hierarchy, was based on a standard 64-kilobits per second (Kbps) signal designed to carry one analog voice signal transformed by pulse-code modulation (PCM) into digital form. This basic unit is known as *DS0*. The International Telecommunication Union (ITU) now supports an entire set of digital signaling standards (Table 1), incorporating elements from the North American (United States/Canada), European, and Japanese standard hierarchies.

The traditional U.S. multiplexing hierarchy began with combining 24 DS0-level signals into one DS1. It is commonly called a *T1* stream, and consists of a sequence of 24 channels combined to create one frame. Each channel is filled with 8 bits (an octet or byte) representing one PCM sample. A particular challenge of the time was to ensure synchronization between transmitter and receiver, which can be accomplished in several ways. For example, each frame could be introduced by a unique starting sequence of 12 bits to allow receiver synchronization to be renewed on a frame by frame basis. The U.S. designers decided instead to distribute the 12 bits over 12 frames, reducing transmission overhead at the expense of receiver complexity. The 12-frame sequence was called a superframe. With improved hardware, synchronization is more easily maintained over longer periods, and an extended superframe (ESF) has replaced the superframe. ESF comprises 24 frames but only needs 6 bits for synchronization, freeing up 4 Kbps that have been used to improve management and control.

In the European scheme (also used by other countries such as Mexico), the basic *E1* stream aggregates 32 PCM channels. Rather than adding synchronization bits, E1 dedicates the first PCM channel for synchronization and the 17th for management and control signaling.

Optical Fiber Systems

Service providers first used digital multiplexing within their own networks (e.g., trunking between Central Offices), to improve the return on and extend the life of their copper cable infrastructure investments. By the 1980s, however, interest had shifted to fiber optics for longer distance, higher speed communications. Standards were defined for the Synchronous Optical Network (SONET in the United States, equivalent to the Synchronous Digital Hierarchy, SDH, in Europe and elsewhere) to carry TDM traffic cost-effectively and reliably over metropolitan and wide area distances. Today SONET specifies both a standard optical interface signal and a digital signaling hierarchy tailored to the fiber transmission environment. The hierarchy is based on an 810-octet frame transmitted every 125 microseconds (μs) to create synchronous transport signal-level 1 (*STS-1*) for electrical signals. Each octet is equivalent to a 64-Kbps PCM channel. For fiber transmission, the STS-1 equivalent is optical carrier-level 1 (*OC-1*). Higher level signals are formed from specific multiples of OC-1 (Table 2). Each SONET frame is structured into transport overhead and a synchronous payload envelope (SPE), which consists of both path overhead and payload. It is only the payload portion that carries subscriber traffic to be routed and delivered through the SONET network.

The major building blocks for SONET networks are the point-to-point multiplexer, and for point-to-multipoint configurations, the add-drop multiplexer (ADM). In particular, the ADM allows traffic to be dropped off and the resultant free capacity to be reused to carry traffic entering the network at that point. SONET ADMs can also be employed to create highly survivable networks that maximize availability using diverse routing and self-healing, survivable ring structures. Figure 3a shows a dual-ring structure where the network accommodates loss of a link

Table 1 Digital Signal Hierarchy

Designation	Capacity (Mbps)	Number of DS0s
DS1	1.544	24
E1	2.048	32
DS2	6.312	96
E2	8.448	128
J3	32.064	
E3	34.368	512
DS3	44.736	672
J4	97.728	
E4	139.264	2,048
DS4	274.176	4,032

DS, North America; E, Europe; J, Japan.

Table 2 Basic SONET Levels

Designation	Line rate	SDH equivalent
OC-1	51.840 Mbps	
OC-3	155.250 Mbps	STM-1
OC-9	466.560 Mbps	STM-3
OC-12	622.080 Mbps	STM-4
OC-18	933.120 Mbps	STM-6
OC-24	1.24416 Gbps	STM-8
OC-36	1.86624 Gbps	STM-12
OC-48	2.48832 Gbps	STM-16
OC-96	4.97664 Gbps	STM-32
OC-192	9.95328 Gbps	STM-64

1 Gbps = 1,000 Mbps; 1 Mbps = 10^6 bits per second.

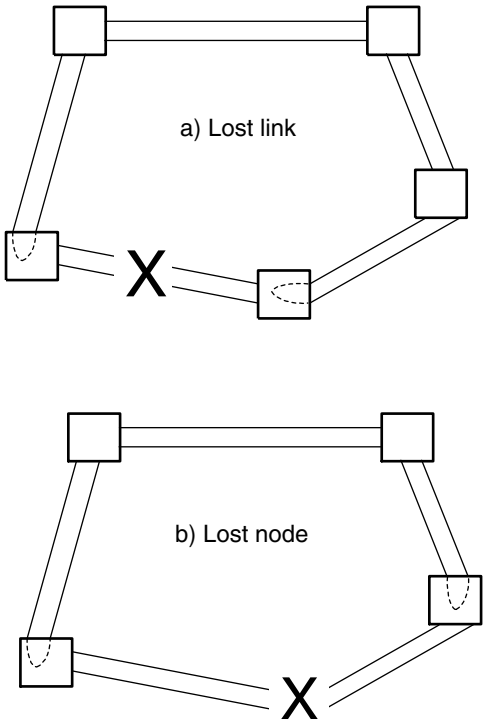


Figure 3: SONET ring survivability, a) lost link and b) lost node.

by looping traffic back on each side of the break, and Figure 3b shows how loss of a network node can be handled similarly. SONET has been deployed extensively by service providers in metropolitan areas to create highly reliable and scalable transport capabilities. Once the fiber and switching equipment are in place, transport capacity can be increased by installing higher-speed signaling interfaces.

Another approach to increasing the capacity of fiber systems has become available with advances in optical component technology. Rather than using the entire range of wavelengths that can be carried over fiber as a single transmission channel, newer equipment allows us to divide the range into multiple channels for simultaneous transmission using wavelength-division multiplexing (WDM). This is quite similar to sending multiple television channels over a coaxial cable. Channels must be spaced far enough apart to limit the interference between adjacent signals that would degrade signal quality. In *coarse* WDM (CWDM) the channels are widely spaced; for *dense* WDM (DWDM), they are very close together (spacing $\leq$ 25–50 GHz). By combining WDM and high-speed signaling, transmission capacities of OC-192, OC-768, and greater become possible, limited primarily by the quality of existing fiber installations.

Access Technologies

In order to get traffic in and out of a MAN or WAN, subscribers must have physical connections, or *access*, to the appropriate service provider's network resources. In the regulated telecommunications environment of the United States, this typically means acquiring connectivity from a LEC to tie the subscriber's physical premises to a WAN service provider's (i.e., IXC's) equipment as physically located in a point of presence (POP). In a metropolitan area, a single company may be allowed to provide both local exchange connections and MAN services. The primary means of accessing MAN and WAN service provider networks are described below.

Dial-up

Dial-up access is appropriate for occasional connections of limited duration, as for making a telephone call. Where the physical facilities used for dial-up were designed and installed to support analog voice traffic, two characteristics are particularly important for data networking:

Digital data must be converted to analog using a modem at the subscriber end and reconverted to digital by a modem at the provider end of the connection.

Data rates are limited by the analog frequency range accepted at provider receiving equipment and by the signal modulation techniques of the modem devices. The most widely accepted standards today support maximum data rates of 56 Kbps.

Leased Line

Leased-line access is more appropriate for connections that need to be continuous and/or of better quality for higher-speed transmission. Such facilities are dedicated to the use of a specific subscriber. For example, a business may lease a T1 access line as its basic unit of connection capacity (1.544 Mbps), particularly for access to Internet service providers. Fractional-T1 and multiple-T1 lines are also available in some areas. A newer technology designed to be digital from end to end over copper cabling, called digital subscriber line (DSL), is beginning to be offered as a lower-cost alternative to the traditional T-carrier. Leased-line access requires matching equipment at each end of the line (subscriber and service provider) to ensure transmission quality suitable to the desired data rates.

Wireless

Wireless access is growing in popularity among mobile individuals who do not work from a fixed desktop in a single building location (e.g., salespeople, customer service representatives, and travelers). Rather than having to find a suitable "land-line" telephone connection with an analog data port to connect the modem, wireless users have either wireless network interface cards or data interface cables that connect their modems to cellular telephones. Both approaches require proximity to a wireless receiving station of matching technology that is then connected to the wired resources making up the remainder of the MAN or WAN.

Cable Modem

Cable modem access is provided by cable television companies who have expanded their business into data networking. A modem designed to transmit data signals over coaxial, broadband television cable is connected, usually via Ethernet technology, to the subscriber's internal network or computer equipment. In residential applications,

subscribers in a neighborhood typically share data networking capacity on the aggregate cable that carries traffic back to the provider's central service location. This is different from a leased-line approach, where access capacity is dedicated from each subscriber location all the way to the provider's POP. In the United States, cable providers are regulated differently from other public telecommunications providers, and may not suffer the same consequences for unavailable service.

Management

Management for MANs and WANs typically began with proprietary systems sold to service providers by each manufacturer of telecommunications switching equipment. Networks composed of equipment from multiple vendors thus contained multiple management systems. Equipment management and service management functions are often tied together by an *Operations Support System* (OSS) in order to automate operations (e.g., performance monitoring), administration (e.g., ordering and billing), maintenance (e.g., diagnostics, fault detection and isolation), and provisioning (OAM&P) functions. Many service providers tailored what they could acquire as a basic OSS in order to accommodate their own specific sets of equipment and services, making it difficult to share information, provide consistent management data in a multiprovider environment, and keep up to date with new functional requirements. This often leaves customers who need services from multiple providers without a single, coherent view of their enterprise WAN resources.

Beginning in 1988, the Telecommunication Standardization sector of the International Telecommunication Union (ITU-T, formerly the Consultative Committee on International Telephony and Telegraphy, CCITT) set about establishing the precepts for a standard *Telecommunications Management Network* (TMN). While the concept of a TMN encompasses the entire set of OAM&P applications in the network, what they do, and how they communicate, ITU-T standards focus on the information required and how it should be communicated rather than how it is processed (M.3000 recommendation series). Two types of telecommunications resources are encompassed: managed systems (such as a switch), which are called network elements (NE), and management systems, usually implemented as operations systems (OS). TMN standards are organized into interface specifications that define the interconnection relationships possible between resources. Figure 4 shows the relationship between the TMN and the telecommunication network for which it is responsible.

TMN is based on the Open Systems Interconnection (OSI) management framework, using object-oriented principles and standard interfaces to define communication for purposes of managing the network. The primary interface specification, Q3, allows direct communication with an OS. Any network component that does not implement Q3 may not access an OS directly, but must go through a mediation device (MD) instead. Legacy equipment and systems that rely on proprietary ASCII messages for communication are accommodated by means of a Q-adapter (QA) that can translate between messages representing the legacy information model and the object-oriented representation expected in today's TMN.

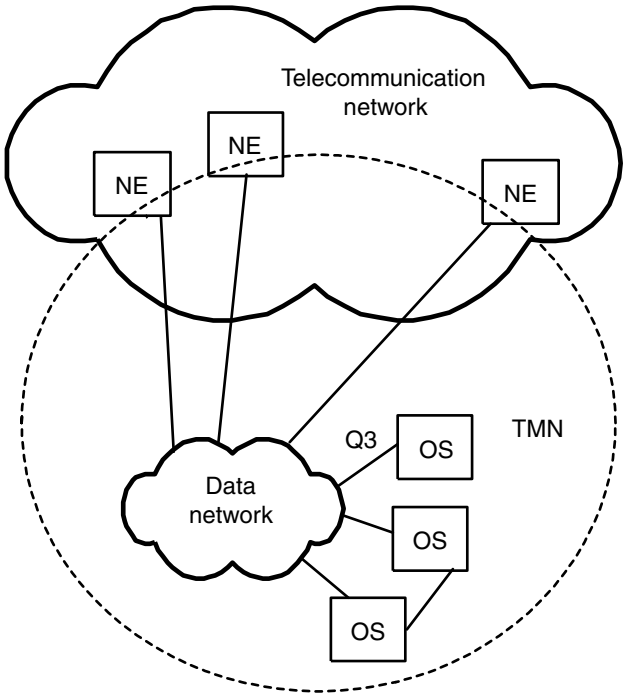


Figure 4: TMN and the network it manages.

TMN defines a layered architecture (ITU-T standard M.3010) as a logical model for the functions involved in managing a telecommunication network effectively (Table 3). The object is to create a framework for interoperability across heterogeneous operation systems and telecommunication networks that is flexible, scalable, reliable, easy to enhance, and ultimately, inexpensive to operate. Standard management services have been defined for alarm surveillance (Q.821), performance management (Q.822), traffic management (Q.823), ISDN service profile management (Q.824), call detail recording (Q.825), and routing management (Q.826).

Differences around the World

Creating and operating WANs or MANs in different countries may present challenges well beyond identifying a service provider and getting connections established. A particular type of service may not be available in the desired location, or a single provider may not offer services in every location, or the capacity required may not be available. Such differences may be due to telecommunication infrastructure of varying ages and technologies, or to different regulations on service offerings in various countries. For example, T1 service is readily available in most U.S. cities. Mexico, however, employs the European standard hierarchy. Thus E1 service would need to be ordered (if it is available) to connect a business location in Mexico to one in the United States, and the differences in capacity and framing would have to be handled appropriately by the network equipment at each end of the link.

In some countries, telecommunication is a regulated industry subject to many government-imposed rules, and

Table 3 TMN Architecture

Logical layer	Functional responsibilities
Business management	Provides an enterprise view that incorporates high-level, business planning and supports setting goals, establishing budgets, tracking financial metrics, and managing resources such as products and people.
Service management	Provides the basic contact point for customers (provisioning, billing and accounting, troubleshooting, quality monitoring, etc.) as well as for service providers and other administrative domains.
Network management	Provides an overall view of the network resources, end to end, based on the information from below about network elements and links. Coordinates activities at the network level and supports the functional requirements of service management.
Element management	Provides a view of individual network elements or groupings into subnetworks. Element managers (OSs) are responsible for subsets of all network elements, from the perspective of TMN-manageable information such as element data, event logs, and activity. Mediation devices belong in this layer, communicating with OSs via the Q3 interface.
Network elements	Presents the TMN-manageable information of individual network resources (e.g., switches, routers, Q-adapters).

there may be no or a limited choice of carriers. Other countries have begun to deregulate, so that multiple carriers compete for subscriber business, often creating more choices in technology and services, as well as better pricing. In either case, service availability may differ from one location to another: DSL access might be easily obtained in greater Boston, but not be available in a rural area; T1 service might be acquired readily in Singapore but perhaps not everywhere in New York City.

Do not make the mistake, however, of assuming that more highly developed areas or countries always have better service options than developing ones. An established metropolitan area experiencing rapid growth in demand for telecommunications may be less able to adapt or expand existing cable and switching capacity to meet new orders than a new suburban business park where there is plenty of room to install new cables and switches to provide higher-speed services. Similarly, developing countries that have very little investment in old infrastructure may be able to skip generations of technology, installing the latest where there was previously none. Economics tend to dictate that this does not happen uniformly, but rather emphasizes locations more likely to provide rapid payback for the particular technology investment (e.g., urban rather than rural, business rather than residential, and high-density population areas). Often it is the access infrastructure that lags behind, because the upgrade costs cannot be amortized across multiple subscribers the way backbone investments can. This is especially true where the end-points are individuals with more limited budgets than business or organizational enterprises.

SWITCHING, ROUTING, AND SIGNALING

Network Architecture

MANs and WANs are usually divided into three logical segments (Figure 5). *Access* typically includes the customer premises equipment (CPE) located in a subscriber's building or office area and the link that physically connects from there to the service provider's point of presence. This link is connected to a device at the *edge* of the service provider's network, and the edge device is connected to devices that compose the *core* (also called the backbone) of the service provider's network. Different technologies are often used in the access and core portions, with the edge required to translate between the two. The ratio of the aggregate input capacity from all subscriber connections to an edge device to the output capacity from the edge into the core describes the degree of oversubscription. For example, if the sum of all access links is 200 Mbps and the core link is 100 Mbps, then the oversubscription ratio is 2:1. A ratio less than or equal to 1 is called nonblocking; the network performance for values greater than 1 depends on the bursty nature of data traffic to minimize the probability that traffic will be delayed excessively (by buffering) or discarded (when buffers become full).

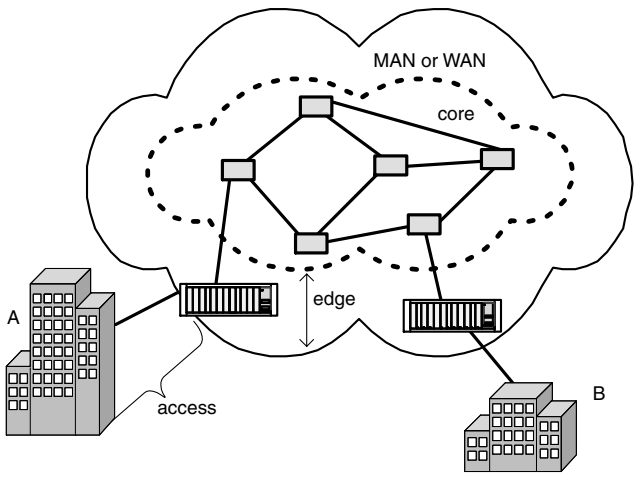


Figure 5: WAN/MAN architecture.

Some form of packet switching is employed in most core data networks today to move traffic through the network. Various techniques are used to meet customer expectations for reliable, timely, and effective delivery of traffic to its intended destination. For example, a *virtual circuit* can be established to approximate the service characteristics available in a circuit-switching environment, such as guaranteed delivery of packets in the same order as they were transmitted. However, virtual circuits do not dedicate resources along the path from source to destination, so the network must have sufficient intelligence to keep traffic moving well enough to meet subscriber expectations.

Choosing the best place to put network intelligence (at the edge or in the core) has been a subject of ongoing discussion among service providers for many years. For example, packets could be examined and labeled at the edge in a way that forwarding decisions in the core are made by simple, high-speed switches. This approach would provide very fast core transit, but the cost of many intelligent edge devices could be high and core switches must still be smart enough to accommodate and adapt to changes in network topology or conditions. An alternative approach makes the edge devices quite simple and inexpensive, while requiring the core to have the intelligence and take the time to understand the characteristics and accommodate the transport needs of the traffic.

Switching Technologies

In the OSI Reference Model, switching takes place at Layer 2, the Data Link Layer. However, much of the WAN switching technology for data networking was developed from experience with X.25, an ITU-T packet-switching protocol standard developed in the 1970s to support public data networking, and still in use today. X.25 creates a connection-oriented network out of packet-switching resources by employing virtual circuits to handle packet flow, keeping the data link layer simpler but requiring circuits to be established before packets can be sent. Circuits that are prebuilt from a source to a particular destination and then left in place are *permanent* virtual circuits (PVCs), while *switched* virtual circuits (SVCs) are established only on demand. SVCs are like dial-up connections, requiring circuit establishment to the specified destination for each call before traffic can flow.

X.25

X.25 is a three-layer protocol suite (Figure 6). The OSI network layer equivalent is the packet-layer protocol (PLP), which has operational modes for call establishment, data transfer, and call termination, plus idle and restarting operations. These functions are implemented through the services of a data link protocol called the Link Access Procedure, Balanced (LAPB), which is responsible for framing data and control commands and for basic error checking through use of a frame-check sequence (Figure 7). During call establishment, the PLP sets up SVCs using X.121 standard addresses. These include the international data number (IDN), made up of a four-digit data network identification code (DNIC, to specify the packet-switching network containing the destination device) and a national terminal number (NTN) consisting of as many as 10 digits. The NTN specifies the exact destination device to which packets will be forwarded.

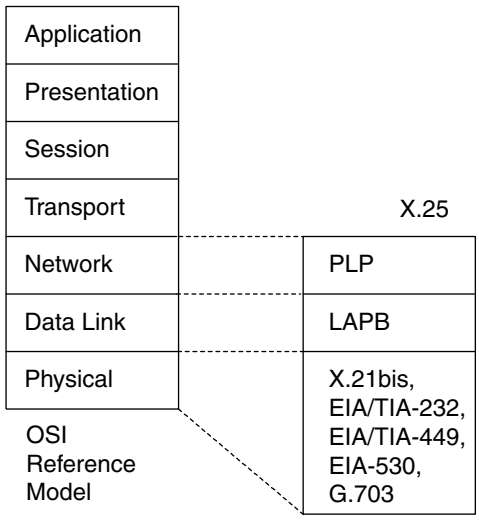


Figure 6: X.25 protocol suite.

Frame Relay

Frame relay is the most widely used packet-switching WAN technology going into the 21st century. As WAN facilities became more reliable during the 1980s, interest rose in streamlining X.25 to improve performance and efficiency. Frame relay (FR) was thus designed as a Layer-2 protocol suite, with work begun by CCITT in 1984. However, it was not until 1991, when several major telecommunication equipment manufacturers formed a consortium called the Frame Relay Forum (FRF) to work out interoperability issues and foster acceptance, that frame relay began to be more widely deployed. In particular, FRF defined extensions to the CCITT work called the local management interface (LMI) to improve service providers' abilities to provision and manage frame relay services.

Frame relay networks (Figure 8) are based on the concepts of data-terminal equipment (DTE) and data circuit-terminating equipment (DCE) first defined by X.25. Subscriber hosts, servers, workstations, personal computers,

Flag (frame delimiter)	Address (command or response indicator)	Control (frame type, sequence #, function)	DATA	FCS (frame check sequence)	Flag (frame delimiter)
1 byte	1 byte	1 byte	variable	2 bytes	1 byte

Figure 7: LAPB frame format.

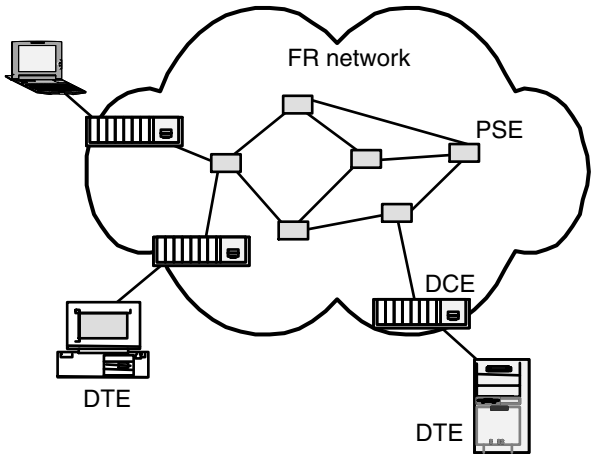


Figure 8: Frame relay network elements.

and terminals connected to a frame relay network are all considered to be DTE. The DCE is usually built as an interface into the service provider's packet-switching equipment (PSE) rather than just being a modem at the edge of an X.25 network. Frame relay also uses virtual circuits to create a bidirectional communication path between a pair of DTE devices. FR virtual circuits are distinguished by data link connection identifiers (DLCIs), which may have local significance only, meaning that each end of a single virtual circuit could have a different DLCI assigned by the FR service provider.

The format for frame relay data combines LAPB's address and control fields into one 16-bit address field that contains the 10-bit DLCI, an extended addressing indicator bit (for future use), a command/response bit that is not used, and congestion control information. To minimize network overhead, the congestion control mechanisms are quite simple:

one forward-explicit congestion notification (FECN) bit that tells a DTE that congestion occurred along the path in the direction *from* the source *to* the destination;

one backward-explicit congestion notification (BECN) bit that tells a DTE that congestion occurred along the path in the direction *opposite* to the transmission from the source to the destination; and

one discard-eligibility (DE) bit to indicate whether this is a lower priority frame that may be discarded before others in a congested situation.

As a packet-switching technology, frame relay also depends on the bursty nature of data traffic to make efficient use of its transmission facilities for larger numbers of subscribers than could be served with physically dedicated connections. The ability to overbook resources is fundamental to the service provider's business model, as well as being a benefit to subscribers, who may be able to insert traffic occasionally at a higher rate than nominal for their access link (called bursting).

Integrated Services Digital Network (ISDN)

Integrated services digital network (ISDN) is a set of telecommunication standards first developed from the perspective of telephony networks to accommodate multiple types of traffic such as voice, fax, data, alarm systems, and video, all in digital format, over a single network. The goal was to develop standard interfaces, both for access and within the network, that would allow all types of digital traffic to be transported end to end, reliably, and in a timely fashion according to the needs of its application. The best-known elements of ISDN are the user interface definitions for connecting subscriber equipment to the network: the primary rate interface (PRI), intended to replace T1 and E1 services, and the basic rate interface (BRI), designed with multiple channels for voice or data traffic from an individual subscriber.

Asynchronous Transfer Mode (ATM)

Asynchronous transfer mode (ATM) was selected as the OSI Layer-2 transport technology for broadband ISDN (B-ISDN) in 1988. It was designed to be useful across WAN, MAN, and LAN communications, as well as to accommodate multiple types of traffic in a single network (voice, data, video, etc.) and scale for very large networks. Other design goals included the abilities to support a variety of media types (e.g., fiber and copper), leverage signaling standards already developed for other technologies, promote low-cost switching implementations (potentially one-tenth the cost of routing), adapt readily to future network requirements, and enable new, large-scale applications. The challenges inherent in such a diverse set of goals brought together designers from many different backgrounds, and resulted in a rather complex architecture (Figure 9).

Basically, ATM is a connection-oriented, packet-switching technology that uses fixed-length packets called *cells*. The 53-byte cell size (5 bytes of header information and 48 bytes for the payload) was chosen as a compromise between the optimal size for voice traffic and the larger size preferred for data applications. The fixed size and format mean that very fast switches can be built across a broad range of transmission rates, from megabits to gigabits per second and beyond. ATM interfaces are often characterized by their equivalent optical-carrier levels whether they employ fiber or copper media. The most

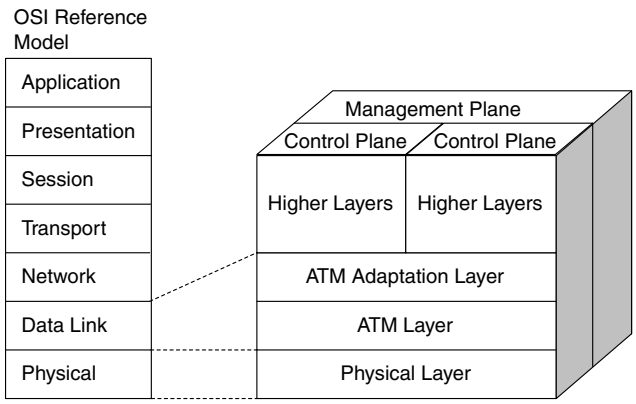


Figure 9: ATM reference model.

popular interfaces tend to be OC-3, OC-12, and OC-48 (Table 2), according to their application in WANs, MANs, or LANs.

An important feature of ATM is the definition of service categories for traffic management:

Constant Bit Rate (CBR) was designed to emulate traditional circuit-switched connections. It is characterized by minimum and maximum cell rates specified at the same, constant value. Typical CBR applications include uncompressed voice and video, or television, all sensitive to both delay and delay variation.

Variable Bit Rate real-time (VBR-rt) and non-real-time (VBR-nrt) are characterized by specified minimum and maximum cell rates, much like frame relay. Typical applications include compressed voice or video, and multimedia e-mail. VBR-rt handles applications sensitive to delay variation, while VBR-nrt is suitable for bursty traffic.

Unspecified Bit Rate (UBR) handles traffic on a best-effort basis, without guaranteeing delivery or any particular rate. This is used to carry data (such as store-and-forward e-mail) not sensitive to delay. In a highly congested network situation, UBR cells may be discarded so that the network can meet its traffic contracts for the other types.

Available Bit Rate (ABR) is characterized by a guaranteed minimum cell rate, but may offer additional bandwidth when network resources are available. Rate-based flow control provides the adjustment mechanism. When it is offered, ABR is often preferred for data traffic.

ATM's service categories are crucial to meeting user demands for *quality of service* (QoS), which generally means guaranteed, timely delivery of traffic to match the needs of particular applications. An ATM end system will request a particular level of service for traffic entering the network, forming a traffic contract with the network. The ATM switches throughout the network are responsible for meeting the terms of the contract by traffic shaping (using queues to smooth out traffic flow) and by traffic policing to enforce the limits of the contract. The capabilities of ATM to provide QoS end to end across a network for multiple types of traffic simultaneously are the most sophisticated to date, and distinguish ATM from other packet-switching technologies. Its suitability for LAN, MAN, and WAN applications makes ATM especially popular with service providers, because they can use one technology throughout to manage their own infrastructure and to support a large variety of service offerings to their customers.

Fiber Distributed Data Interface (FDDI)

Fiber distributed data interface (FDDI) was developed by the American National Standards Institute (ANSI) in the mid-1980s as a 100-Mbps standard for ring-based networks that had outgrown their capacity to handle high-speed workstations or provide nonblocking backbone connections. It was designed originally to expand the typical LAN environment, using a timed token access method for sharing bandwidth at OSI Layer 2 and readily available optical components and fibers for the Physical Layer. Ring management functions are distributed, with single- and dual-ring topologies supported, to create a highly reliable network with deterministic, predictable performance. FDDI was the first LAN technology suitable for distances beyond a building or small campus, and was used by some to cover the geographic scope of a MAN.

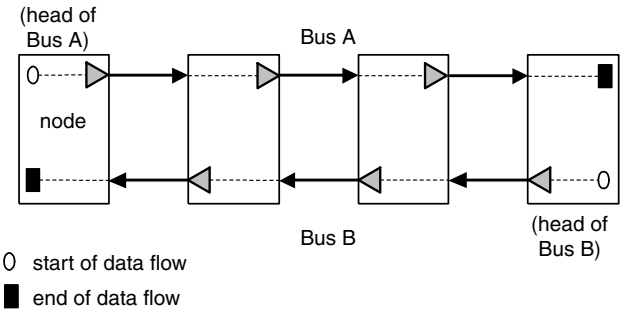


Figure 10: DQDB architecture (ANSI/IEEE Std 802.6, 1994 edition).

Distributed Queue Dual Bus (DQDB)

Distributed queue dual bus (DQDB) was also developed during the 1980s, specifically to address the needs of metropolitan area networking for integrated services such as voice, data, and video. The IEEE 802.6 working group finally ratified it as a Layer-2 standard in 1990. As its name suggests, DQDB specifies a network topology of two unidirectional buses that are able to interconnect multiple nodes (Figure 10). The supporting physical layer for DQDB initially offered various transmission interfaces and speeds from DS3 (45 Mbps) to STM-1 (155 Mbps). The idea of DQDB was that multiple subnetworks could be interconnected to form a MAN, with the goal of supporting connectionless and connection-oriented data transfers, along with isochronous traffic, sharing the total communication capacity available.

DQDB may be most familiar as the basis for definition of *switched multimegabit data service* (SMDS) packet-switched public data networks. SMDS was designed by Bell Communications Research (Bellcore) for high-speed, connectionless delivery of data beyond the LAN. Its variable frame size up to 9188 octets is large enough to encompass as payload any of the popular LAN technology frames (i.e., Ethernet, token ring, and FDDI). The SMDS interface protocol was defined as a three-level protocol that specifies how subscribers access the network. As a service, SMDS was intended to be independent from any underlying transport technology. Thus it was first offered at DS1 to DS3 access speeds, with a goal of increasing later to OC-3.

Ethernet

Ethernet became the dominant LAN technology in the latter 1990s, as extensions from the original 10 Mbps were defined for 100 Mbps, then 1,000 Mbps (= 1 Gbps), and became widely deployed. In the same time period, new communication companies with no telephony heritage began laying optical fiber, and leasing capacity for short-haul (i.e., MAN) or long-haul (i.e., WAN) connections rather

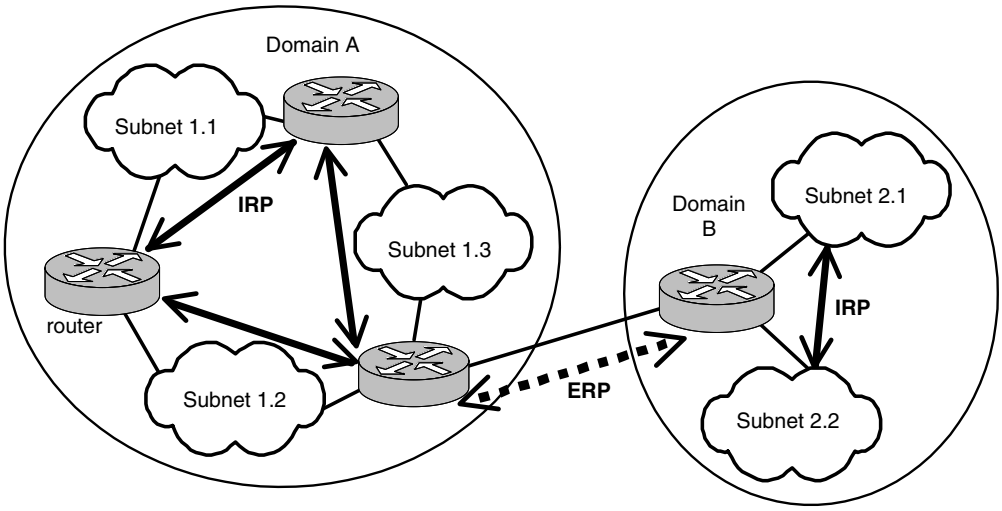


Figure 11: Routing within and between autonomous domains.

than selling services, as was typical in public networks. This meant that customers could specify the technology used to put bits on the medium rather than subscribing only to specific services offered by providers. As advances in optics and use of switching allowed Ethernet to cover even greater distances, the geographic limits that distinguished LAN from MAN technologies began to disappear. In fact, new providers sprang up offering Ethernet connectivity from the business doorstep to other locations across town or beyond. The great competitive question was whether Ethernet MANs could be made as reliable and fault-tolerant as more traditional MAN/WAN technologies built over SONET.

Resilient Packet Ring (RPR)

Resilient packet ring (RPR) is an effort begun by the IEEE 802.17 working group in late 2000 to design a high-speed access protocol combining familiar Ethernet interfaces with the fault-tolerance and rapid restoration capability of ring-based MAN technologies like SONET. RPR defines a new medium access control (MAC sublayer of OSI Layer 2) protocol that extends Ethernet framing from the LAN into the MAN/WAN environment. As seen by the RPR Alliance (an industry consortium designed to promote adoption of RPR), this approach combines the cost-effective scalability of Ethernet access interfaces with a MAN that can be optimized for rapidly increasing volumes of data traffic. Because it focuses on the MAC sublayer, RPR is independent of the underlying Layer-1 technology, making it suitable to run over much of the MAN infrastructure already in place.

Routing Technologies

In the OSI Reference Model, routing takes place at Layer 3, the Network Layer. Essentially routing consists of three major functions: maintaining information about the network environment, finding a path through the network from particular sources to destinations, and forwarding packets at each relay point. The Internet protocol (IP) is the dominant method of interconnecting packet-switched networks (i.e., for internetworking) at Layer 3. It provides connectionless network services (CLNS), with no guarantee of delivery or packet ordering, and is widely used today for private and public LANs, MANs, and WANs, including the Internet. IP is primarily concerned with the format for packets (also called datagrams), the definition and structure of addresses, a packet-forwarding algorithm, and the mechanisms for exchanging information about conditions in and control of the network.

Routing responsibility in an internetwork is divided between intradomain or interior routing protocols (IRPs) and interdomain or exterior routing protocols (ERPs) as shown in Figure 11. IRPs are used for internetworks that belong to a single administrative authority, such as an enterprise LAN, a single service provider's MAN, or a private WAN. ERPs are used when routers tie together networks belonging to multiple independent authorities, as in the Internet. These protocols differ in how much information is kept about the state of the network and how routing updates are performed using the mechanisms defined by IP.

IP Version 4 (IPv4)

IP version 4 (IPv4) was defined by the Internet Engineering Task Force (IETF) for the original ARPAnet and published as (Request for Comments) RFC 791 in 1981. It specifies that each interface capable of originating or receiving internetwork traffic be identified by a unique 32-bit address consisting of an ordered pair containing a network identifier (net_ID) and a host/interface identifier (host_ID). Three primary classes of network addresses (A, B, and C) were designed to promote efficient routing, with additional classes defined for special or future uses (Figure 12). Although the Internet is not centrally managed, it was necessary to establish a single authority to assign addresses so that there would be no duplicates or conflicts.

As the Internet grew through the 1980s, a number of limitations in the design of IPv4 became apparent. The allocation of addresses, especially classes A and B, tended to be wasteful. For example, a single class B address assigned to one organization accommodates one network with over 64,000 IP interfaces—much larger than is practical or

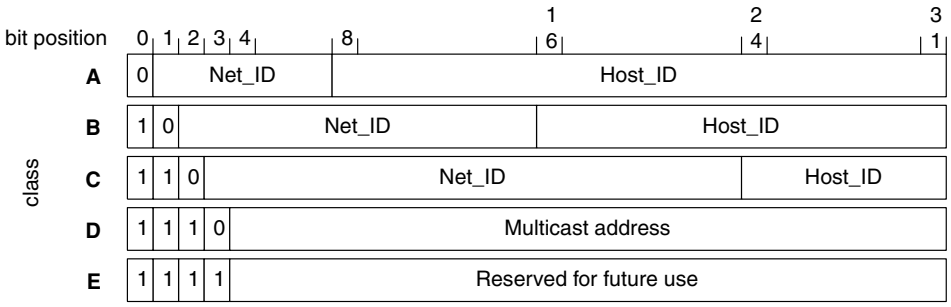


Figure 12: IPv4 addressing format.

needed for most, meaning that a lot of address space can be wasted. On the other hand, a single class C address accommodates only 255 interfaces, which is too small for most organizations, requiring them to have more than 1. From a routing perspective, the two-level hierarchical address structure means that routers need to keep track of over 16 million net_IDs just for class C networks, as well as calculate paths through the Internet to each one. A number of schemes were developed to solve some of the addressing and router problems (subnet masking, classless interdomain routing or CIDR), but those were not the only issues. Rising interest in using the Internet to carry voice, video, multimedia application, and commercial transaction traffic increased the demand for security and quality of service support, neither of which were built into IPv4. Consequently, the IETF began work on a new version, IP-ng, to handle the next generation.

IP Version 6 (IPv6)

IP version 6 (IPv6) represents that next generation of Network Layer services. It extends the addressing space from 32 to 128 bits, simplifies the packet header and allows for future expansion, and adds new capabilities to label flows of packets (same source to a single destination), to assign packets priority in support of QoS handling, and to provide authentication and security. Several of these features (CIDR, DiffServ, and IPsec) were designed so they could be added onto IPv4. In fact, such retrofitting solved IPv4 problems well enough in the late 1990s that people began to question whether a move to IPv6 was necessary. Upgrading the large numbers of routers involved with Internet traffic would be expensive, time-consuming, and require careful coordination. Transition strategies and mechanisms would likely be needed over a considerable period of time. Unfortunately, retrofits cannot do much about the size of IPv4 addresses. Sufficient growth in the numbers and types of devices people want to connect to or through the Internet (handheld devices, household appliances, automobile systems, etc.) and international pressure from countries without enough addresses will eventually make IPv4 addressing inadequate. The only question seems to be when.

Border Gateway Protocol (BGP)

Border gateway protocol (BGP) is the exterior routing protocol used by independent or autonomous systems (ASs) to exchange routing information throughout the Internet. Published in 1995 as RFC 1771, it defines procedures to establish *neighbor* relationships, and to test the *reachability* of neighbors and other networks. A router at the edge of an AS uses BGP to work with adjacent (i.e., directly connected) routers in other ASs. Only after two routers (one in each AS) have agreed to become neighbors can they exchange routing information or relay traffic for each other's AS. Unlike IRPs, which use the services of IP to accomplish their communication, BGP uses the reliable transport services of TCP (transmission control protocol, running over IP). In this way, BGP can be simpler because it depends on the error control functions of TCP, and its messages are not limited in size by the constraints of an IP datagram.

BGP is purposefully designed to allow an AS to control what detail of internal information is made visible outside the AS (aggregating routes using CIDR, for example). Typically each BGP router screens potential routing updates or reachability advertisements against a configuration file that specifies what type of information it is allowed to send to each particular neighbor. This approach promotes policy-based routing, but at the expense of needing to calculate paths from incomplete detail about the network topology. Thus BGP will not always choose the optimal path across an internetwork to reach a particular destination. It does, however, allow a country or company constituting an AS to make appropriate political or business decisions about when and where to route its traffic.

Questions about the scalability of BGP have been raised in light of predictions for continued substantial growth in Internet traffic, and particularly as more organizations consider deploying delay-sensitive applications over the Internet (e.g., voice, video, conferencing). Intelligent route control, virtual routing, and new approaches to traffic engineering are among the options being explored to solve performance problems before they become serious impediments to effective use of the Internet.

Multiprotocol Label Switching (MPLS)

Multiprotocol label switching (MPLS) has been designed by the IETF to improve the performance of routed networks by layering a connection-oriented framework over an IP-based internetwork. MPLS requires edge routers to assign labels to traffic entering the network so that intermediate routers (called label-switching routers, LSRs) can make forwarding decisions quickly, choosing the appropriate output port according to the packet's label and rewriting that label (which is intended to have local

significance only) as necessary. MPLS represents a significant shortcut from the usual IP approach, where every relay node must look deeply into the packet header, search a routing table for the best match, and then select the best next hop toward the packet's destination. All packets with the same MPLS label will follow the same route through the network. In fact, MPLS is designed so that it can explicitly and flexibly allocate network resources to meet particular objectives such as assigning the fastest routes for delay-sensitive packet flows, underutilized routes to balance traffic better, or multiple routes between the same end-points for flows with different requirements. This is called *traffic engineering* and serves as the foundation for both optimizing performance and supporting QoS guarantees.

Nothing about the MPLS design limits its use to the IP environment; it can work with suitably equipped ATM and frame relay routers as well. In fact, it can coexist with legacy routers not yet updated with MPLS capabilities, and it can be used in an internetwork that contains a mix of IP, ATM, and frame relay. Another powerful feature is the ability to stack labels on a last-in-first-out basis, with labels added or removed from the stack by each LSR as appropriate. This allows multiple label-switched paths to be aggregated into a tunnel over the common portion of their route for optimal switching and transport. MPLS is also a convenient mechanism to support virtual private networks, especially when multiple Internet service providers are involved along the path from one end to the other.

Signaling and Interworking

Connection-oriented networks require specific mechanisms for establishing a circuit (physical or virtual) prior to traffic flow, and for terminating the circuit afterward. In the circuit-switched telephony environment, call setup and termination are part of a well-developed set of telecommunication system control functions referred to as *signaling*. MANs and WANs that were built for voice included signaling as an integral part of their designs, because resources were dedicated to each call as it was established and needed to be released after call completion.

The ITU-T began developing standards for digital telecommunication signaling in the mid-1960s; these have evolved into common channel interoffice signaling system 7 (CCIS7, known in the United States as Signaling System 7, or just *SS7* for short), currently in use around the world. SS7 is an out-of-band mechanism, meaning that its messages do not travel across the same network resources as the conversations it was designed to establish and control. In fact, SS7 uses packet switching to deliver control messages and exchange data, not just for call setup, but also for special features such as looking up a toll-free number in a database to find out its real destination address, call tracing, and credit card approvals. Out-of-band delivery of the messages allows SS7 to be very fast in setting up calls, to avoid any congestion in the transport network, and also to provide signaling any time during a call.

The SS7 network has a number of elements that work together to accomplish its functions (Figure 13):

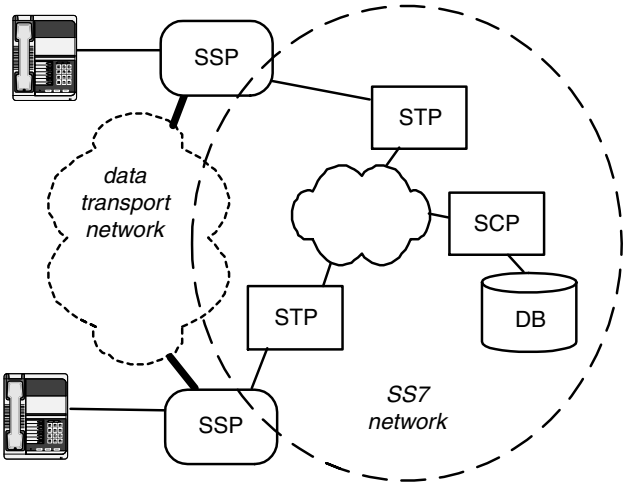


Figure 13: SS7 network elements.

Signal switching points (SSPs) are the network edge devices responsible for setting up, switching, and terminating calls on behalf of connected subscriber devices, and thus insert user traffic into, and remove it from, the service provider's backbone network.

Signal transfer points (STPs) are packet switches responsible for getting SS7 messages routed through the control network.

Signal control points (SCPs) house the databases that support advanced call processing.

In packet-switched MANs and WANs, signaling had been associated primarily with establishing and tearing down SVCs that required no further control during the data transfer phase. With a rising interest in multimedia communications (e.g., video, and especially voice over IP) however, the ITU-T quickly recognized a need for additional capabilities. Their *H*.323 recommendations encompass an entire suite of protocols that cover all aspects of getting real-time audio and video signals into packet form, signaling for call control, and negotiation to ensure compatibility among sources, destinations, and the network. H.323 takes advantage of prior ITU work (such as ISDN's Q.931 signaling protocol) and defines four major elements (Figure 14):

Terminals are the end-user devices that originate and receive multimedia traffic.

Gateways primarily handle protocol conversions for participating non-H.323 terminals, as would be found in the public switched telephone network (PSTN).

Gatekeepers are responsible for address translation, call control services, and bandwidth management.

Multipoint Control Units (MCUs) provide multiconferencing among three or more terminals and gateways.

The IETF took a simpler approach to signaling with the *session initiation protocol* (SIP), which was designed as a lightweight protocol simply to initiate sessions between users. SIP borrows a great deal from the hypertext transfer protocol (HTTP), using many of the same header fields,

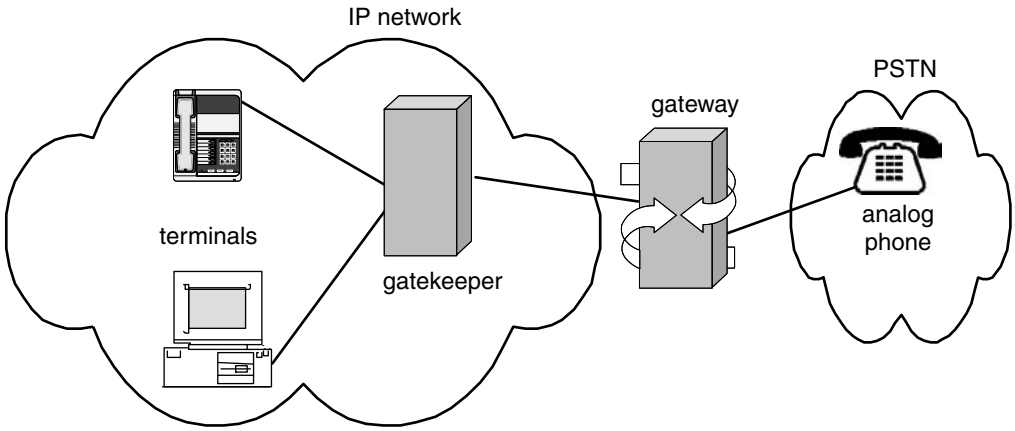


Figure 14: H.323 network elements.

encoding rules, error codes, and authentication methods to exchange text messages. Like H.323, SIP assumes that the end-point devices (i.e., terminals) are intelligent, running software known as the user agent. The agent has two components: *User Agent Client*, which is responsible for initiating all outgoing calls, and the *User Agent Server*, which answers incoming calls. In the network itself, SIP provides support with three types of server:

Registration servers keep track of where all users are located.

Proxy servers receive requests and forward them along to the next appropriate hop in the network.

Redirect servers also receive requests and determine the next hop, but rather than forwarding the request, they return the next-hop server address to the requester.

An alternative approach to multimedia communication control developed by the IETF is called the media gateway control protocol (MGCP). It is quite different from H.323 and SIP because it assumes that the end-user devices are not very intelligent. Consequently MGCP takes a central server approach to communication coordination and control. Two elements are defined: the *Media Gateway Controller* (also known as the call agent), which provides the central intelligence and controls all of the *Media Gateways*, which perform a variety of interface functions such as with the PSTN, residential devices, and business private branch exchanges (PBXs). MGCP defines the communication that takes place between the call agent and the Gateways that execute its commands.

In practice H.323, SIP, and MGCP will likely coexist to support multimedia communication in the Internet environment because each has advantages for specific applications or coverage. MGCP is particularly useful to MAN/WAN service providers with large installed bases of unintelligent end-point devices, and its gateway approach allows for tailored interfaces to each different underlying technology. The simplicity of SIP is more attractive to enterprise networks designed primarily for data traffic with smaller requirements for supporting voice and video. Finally, H.323 is the most mature and most comprehensive. As usual in the telecommunication industry, vendor support and suitability to customer business models are likely to determine which, if any, one approach becomes dominant.

PROVIDERS AND SERVICES

Carriers and Service Providers

The public provision of telecommunication services to subscribers for a fee has a history of being government-regulated in most parts of the world (the term "common carrier," for example, dates back to public transportation for people, first by stagecoach, then by trains, buses, etc.). Regulation was required because access to telecommunication services depended on cabling that was run from subscriber premises (residential or business) across public property (e.g., along roads) to a provider's central office as a service point. Governments could also impose standards to ensure that services offered by providers in different locations would be compatible enough to interoperate. In some countries, infrastructure was built and services operated by the government itself (e.g., PTTs that provided postal, telegraph, and telephone services nationwide). In the United States, telephone industry regulation was divided between LECs whose cabling and local services go to individual premises, and IXCs who provided the interconnection (i.e., long-distance services) between LECs.

The Internet as a means of public data communication has grown up rather differently, driven largely by the U.S. regulatory environment, where telecommunication companies were prohibited from providing data services. Consequently, a new type of company called an Internet service provider (ISP) was born. Data would move from a subscriber's premises, across cables belonging to an LEC, to ISP equipment in a point of presence, where it was transferred onto Internet resources. The subscriber thus had to be a customer of both the LEC and the ISP unless a private link could be installed directly to the ISP's POP. The Internet connections from one ISP location to another are most often lines leased from an IXC. As telecommunication services have been increasingly deregulated world-wide, the distinctions among voice and data service providers have become blurred.

It is important to remember that "the Internet" is not really a single entity, but rather an interconnected set of autonomous networks whose owners have agreed to cooperate and use a common set of standards to ensure interoperability. *Peering* is a form of interconnection where ISPs agree to exchange traffic for their respective customers, based on a specific set of business terms. Peering points are where the networks actually connect to effect this exchange. The number and location of peering points and partners is decided by each ISP according to customer demand and its own business criteria. Subscribers may need to be aware of these agreements in order to understand fully the performance they can expect end to end across the Internet.

Just as the background and emphasis of traditional voice and traditional data service providers differ, so do their business models and their choices of technology. Some offer only transport for traffic, either between subscriber sites or to the Internet. Others offer access to applications or management services. Local telecommunication carriers tend to offer MAN services over an ATM and SONET infrastructure, while data providers would be more likely to offer IP services or simply Ethernet access and transport. Cable television and wireless service providers also offer access services according to the characteristics of their infrastructure technologies. The options available will likely continue to grow as technology progresses.

Class of Service, Quality of Service

As interest in carrying multimedia or multiple-service traffic (i.e., voice, data, video) over MANs and WANs has grown, managing the traffic to provide performance appropriate to each application has become more important. Quality of service techniques are expected to guarantee performance and delivery, usually in terms of bandwidth allocation, timeliness of delivery, and minimal variation in delay (e.g., ATM service categories). Class of service (CoS) techniques do not make such guarantees, but rather attempt to meet user requests on a best-effort basis. Typically CoS works by grouping together traffic with similar requirements (e.g., voice or streaming video) and using a priority queuing system so that switches and routers forward the traffic accordingly. Connectionless network services such as IP offer CoS traffic management, while connection-oriented services such as ATM provide QoS.

QoS cannot really be guaranteed unless it is available all the way from end to end of the connection. This creates a challenge for MAN and WAN environments where multiple technologies from one or more service providers may be involved in delivering user traffic, and especially when the traffic originates or terminates in a LAN of yet another different technology. Several groups are involved in developing standard techniques for CoS and QoS. The problem is making sure that appropriate translation mechanisms can carry user application requirements across network and SP boundaries:

IEEE 802.1p is a Layer-2 tagging mechanism to specify priority using 3 bits in the Layer-2 frame header.

IETF's differentiated services (DiffServ) indicates how packets are to be forwarded using per-hop behavior (PHB) queuing, or discarded if there is not sufficient bandwidth to meet performance requirements.

ATM traffic management defines service categories and traffic classes.

Virtual Private Networks

A virtual private network (VPN) is a special service that amounts to establishing a closed user group capability over a shared or public network infrastructure. This means that access is restricted to authorized users only, privacy of data content is assured, traffic belonging within the VPN does not get out or become visible to unauthorized users, and outside traffic does not get in. VPNs are becoming a very attractive way for organizations to reduce the cost of private WANs while improving the security for traffic that travels over public networks. Where high-speed MAN and WAN services are available, long-distance performance can even be kept reasonably close to what the remote users would experience if they were directly connected to the LAN. VPNs may also be built to send traffic across the Internet, with one or more SPs providing the access links between the Internet and various geographically dispersed customer sites. Internet VPNs can be significantly less expensive than the private lines or networks they replace.

Management

The OSI model for network management encompasses five functional areas: configuration management, performance management, fault management, accounting management, and security management. A MAN or WAN service provider must cover these from the perspective of both operating the entire network effectively and balancing the needs and expectations of paying customers who could always choose to take their business elsewhere. Operation must be reliable, there must be sufficient capacity to meet traffic needs and performance expectations, and privacy must be maintained not only for the content of the traffic carried but also for data about the customers. At the same time, subscribers typically want the ability to manage the performance and flow of their own traffic through their allotment of SP resources. SP operation systems must be capable and sophisticated to meet all these requirements.

A primary mechanism used to establish and manage expectations between customers and providers is the *service level agreement* (SLA). SLAs are the defining documents (contracts) that spell out what services and levels of support will be provided to the customer at a specified price. Successful SLAs are built on a solid, shared understanding of business priorities and service impact, for both the service user and the service provider. Detail about roles and responsibilities, metrics and reporting, added cost for incremental services or enhancements, escalation procedures, and change management are just some of what should be covered in an SLA. Many customers also build in penalties in case the provider fails to deliver services at the level specified in the SLA. This may be

necessary legally to protect the service user's interests, but it is important to remember that failure of the provider to deliver service typically means that the user has failed to meet a business requirement as well. Consequently it is in both the customer's and provider's best interests if the penalty clause is never invoked.

CONCLUSION

The boundaries between local area, metropolitan area, and wide area networks have become less clear over time, due to the increasing variety of implementation choices available to network designers. For MANs and WANs in particular, classifying equipment, services, and management responsibilities by architectural category (access, edge, and core) may help to distinguish among options so that choices can be made in alignment with the business priorities of each designer's organization. Flexibility to accommodate change and growth, reliable service delivery, and optimal cost/performance ratios are the major characteristics typically desired for every network design. The tension between network users wishing to maximize cost-effectiveness and service providers trying to maximize profit continues to work with technological developments to create opportunity for new business approaches. The primary challenge always seems to lie in balancing expensive, long-term infrastructure investments with new technologies and services that meet changing application requirements.

GLOSSARY

Circuit A complete connection between a source and destination for the purposes of transferring information, i.e., for communication.

Circuit switching A *connection-oriented* approach to networking where the entire path from source to destination is determined and sufficient resources are allocated along the path to carry the traffic before any traffic can flow.

Local area network (LAN) A network covering a single office, building, or campus that is built and operated as a private network, by an individual or organization, for his/her/its own use.

Metropolitan area network (MAN) A network covering distances up to tens of miles, often within or surrounding a major city, with public ones built and operated by service providers who offer network services to subscribers for a fee.

Packet A portion of a message to be carried from a source to a destination.

Packet switching A *connectionless* networking approach where each packet is routed through the network independently.

Private network A network built, owned, and operated by a single individual or organization for his/her/its own use.

Public network A network built to offer resources or services to a set of subscribers who are typically independent from each other and from the owner of the network. Most people think of the Internet as the only truly "public" network.

Routing Determining where a packet should go next to get it closer to its intended destination, i.e., deciding what is the next hop along the path.

Service provider (SP) Builder and/or operator of a network for the purpose of selling capacity or services to subscribers for a fee.

Service user Subscriber to the offerings of a public network.

Switching Placing a packet on the appropriate transport mechanism to get it to the network device representing the next hop.

Wide area network (WAN) A network covering distances up to hundreds or thousands of miles, such as between cities, across or between countries, or across oceans, where public facilities are built and operated by service providers who offer network capacity or services to subscribers for a fee while private ones are built by organizations for their own use.

CROSS REFERENCES

See *Circuit, Message, and Packet Switching; Conducted Communications Media; Local Area Networks; Propagation Characteristics of Wireless Channels; Public Networks*.

FURTHER READING

ATM Forum (2002). ATM service categories: The benefits to the user. Retrieved September 17, 2002, from http://www.atmforum.com/aboutatm/6.html

Cisco Systems, Inc. (2001). Fundamentals of DWDM technology. Retrieved September 17, 2002, from http://www.cisco.com/univercd/cc/td/doc/product/mels/cm1500/ dwdm/dwdm_ovr.htm

Cisco Systems, Inc. (2002). Switched multimegabit data service. Retrieved September 17, 2002, from http://www.cisco.com/univercd/cc/td/doc/cisintwk/ito_doc/smds.htm

Defense Information Systems Agency (n.d.). ANSI T1 standards related to TMN. Retrieved September 17, 2002, from http://www-comm.itsi.disa.mil/tmn/tmn_t1.html

DSL Life (n.d.) DSL tutorial. Retrieved September 17, 2002, from http://www.dsllife.com/dsltut.htm

Frame Relay Forum (2002). Retrieved September 17, 2002, from http://www.frforum.com

H.323 Forum (2002). Retrieved September 17, 2002, from http://www.h323forum.org

Horak, R. (n.d.) T-carrier basics. Retrieved September 17, 2002, from http://www.commweb.com/techcenters/main/experts/3783/COM20010726S0011

Welcome to get IEEE 802 (n.d.). Retrieved September 17, 2002, from http://standards.ieee.org/getieee802/

International Engineering Consortium (2002). Cable modems. Retrieved September 17, 2002, from http://www.iec.org/online/tutorials/cable_mod

International Engineering Consortium (n.d.). Various other tutorials. Retrieved September 17, 2002, from http://www.iec.org/online/tutorials

International Telecommunication Union, Telecommunication Standardization Sector (n.d.) Retrieved September 17, 2002, from http://www.itu.int/ITU-T/

Internet Engineering Task Force (n.d.a) Retrieved September 17, 2002, from http://www.ietf.org

Internet Engineering Task Force (n.d.b) Differentiated services (diffserv). Retrieved September 17, 2002, from http://www.ietf.org/html.charters/diffserv-charter.html

IPv6.org (n.d.) IPv6. Retrieved September 17, 2002, from http://www.ipv6.org

MPLS Forum (n.d.) Retrieved September 17, 2002, from http://www.mplsforum.org

OpenH323 (2002). *H.323 standards*. Retrieved September 17, 2002, from http://www.openh323.org/standards.html

Pacific Bell Internet (1999). *Classless Inter-Domain Routing (CIDR) overview*. Retrieved September 17, 2002, from http://public.pacbell.net/dedicated/cidr. html

Performance Technologies (2002). *SS7 tutorial*. Retrieved September 17, 2002, from http://www.pt.com/tutorials/ss7

Protocols.com (n.d.). *ISDN*. Retrieved September 17, 2002, http://www.protocols.com/pbook/isdn.htm

Resilient Packet Ring Alliance (2001, October). *An introduction to resilient packet ring technology*. Retrieved March 16, 2003, from http://www.rpralliance.org/articles/ACF16.pdf

RFC Editor (2002, August 12). Retrieved September 17, 2002, from http://www.rfc-editor.org/

Rybczynski, T. (1999). *IP: Getting some class*. Retrieved September 17, 2002, from http://www.nortelnetworks.com / solutions / financial / collateral / nov98_ipcos_v1.pdf

Sangoma Technologies (n.d.). *X.25 tutorial*. Retrieved September 17, 2002, from http://www.sangoma.com/x25.htm

SIP Forum (2002). Retrieved September 17, 2002, from http://www.sipforum.org

SONET.com (2000, April 26). *Educational information*. Retrieved September 17, 2002, from http://www.sonet.com/edu/edu.htm

Spurgeon, C. (2002, August 24). *Charles Spurgeon's Ethernet Web SITE*. Retrieved September 17, 2002, from http://www.ethermanage.com/ethernet/ethernet.html

University of New Hampshire InterOperability Lab (1998, July 27). *802.1p/Q VLANs*. Retrieved September 17, 2002, from http://www.iol.unh.edu/menu/training/vlan.html

University of New Hampshire InterOperability Lab (1997, July 28). *FDDI tutorials and resources*. Retrieved September 17, 2002, from http://www.iol.unh.edu/training/fddi/htmls

Web Host Industry Review, Inc. (2002). (VPN) Virtual Private Network News. Retrieved March 16, 2003, from http://findvpn.com/news/

Windows 2000 Security

E. Eugene Schultz, *University of California–Berkley Lab*

WHAT IS W2K?

Microsoft's Windows 2000 (W2K) is an operating system product that includes both workstation (Windows 2000 Professional) and server (such as Windows 2000 Server and Windows 2000 Advanced Server) versions. It supports not only desktop and office automation applications but can also be used to run network applications that support mail, Web, and file transfer services, a domain name service (DNS) server, and even routing and firewalling network traffic. W2K also includes many features that were not available in W2K's predecessor, Windows NT (NT), the most notable of which is W2K directory services (called Active Directory). Active Directory provides an infrastructure and related services that enable users and applications both to locate and access objects such as files and printers and services throughout the network. Active Directory is a directory service (similar to Novell's Netware Directory Service) that acts as the main basis for holding and distributing data about accounts, groups, Organizational Units (OUs), security policies, services, domains, trust, and even Active Directory itself. This directory service not only stores data of this nature but also makes it available to users and programs, providing updates as needed.

Active Directory also supports security by storing security-related parameters and data and supporting services (e.g., time services) needed for achieving system and network security. Active Directory is, in fact in many respects, the "center of the universe" in W2K.

HOW W2K WORKS

A good starting point in exploring how W2K works is W2K domains, the focus of the next part of this chapter.

Domains

W2K machines can be configured in either of two ways: as part of a domain or as part of a workgroup consisting of one or more machines. A domain is a group of servers and (normally) workstations that are part of one unit of management. Each domain has its own security policy settings. Policies are rules that affect how features and capabilities in W2K work; they can determine allowable parameters (such as the minimum number of characters in passwords), enable functions (such as the right to increase or decrease the priority with which a program runs), or restrict the ability to perform these functions (I cover policies in more detail later in the chapter). Domain controllers (DCs) hold information related to policies, authentication, and other variables. When a change to a policy is made, a new account is created or deleted, or a new OU is created, the changes are recorded by a DC within a domain, and then replicated to all the other DCs within the domain within a designated time interval.

Domains are good for security, provided, of course, that they are set up and maintained properly. It is possible to set domain policies so that (with a few exceptions) they will be applied to virtually every server and workstation within a domain. This decreases the likelihood that any system within the domain will be a "weak link" system, one that is an easy target for attackers. Additionally, domain functionality includes important features such as the ability limit the workstations and servers that may be added to a domain.

The other option is to belong to a "workgroup." By default, a system that is not part of a domain is a member of its own workgroup. In workgroups, anyone with Administrator privileges on a workstation or server and who knows the name of a certain workgroup can add that machine to the workgroup, something that makes possible discovering a great deal of information about each machine and user in the workgroup. This information can be used advantageously to attack the other systems. Access to resources (such as files, folders directories, printers, and so forth) is determined locally by the particular server or workstation within the workgroup that contains the resources. No built-in central control capabilities exist. Users whose machines are part of workgroups can engage in functions such as sending mail, transferring files,

and so forth, but workgroups are not at all conducive to security because there is no mechanism within W2K to limit workgroup membership. If an attacker discovers the name of a workgroup, that person can add a malicious system to it. Additionally, the lack of centralized control in a workgroup necessitates setting security parameters and adjusting configurations on every machine within the workgroup; in contrast, domains have settings (embedded in "Group Policy Objects" or GPOs) that can be set from a single domain controller (to be defined shortly) within the domain.

The following sections consider various possible relationships between domains and the implications of each.

Trees and Forests

W2K domains can be arranged in a hierarchical fashion starting with a root domain at the top, then domains at the level immediately below the root domain, and then possibly still other domains at the next level(s). One option is to nest domains so that they form a "contiguous namespace." In simple terms, this means that there is one common root domain; all subordinate (lower) domains' names are derived from their parent domains. Consider the name of one domain, research.entity.org. Consider also marketing.entity.org. If the domains are nested in a contiguous name space, both of the domains in this example will have the same parent domain, entity.org. If research.entity.org is a parent domain, every one of its children will have a first name followed by research.entity.org (see Figure 1 below). Contiguous name spaces characterize W2K "trees." In contrast, if the name space is not contiguous, then there is no common namespace. "Forests" (as opposed to "trees") are characterized by noncontiguous name spaces. In a tree or forest, every domain connected directly to another domain (as are entity.org and research.entity.org) by default has a two-way trust relationship with every other domain. Note that if domains are not directly connected to each other (as in the case of marketing.entity.org and research.entity.org in Figure 1), they nevertheless have transitive trust between them because entity.org has a two-way trust relationship with each of its child domains. Trust is a property that allows users, groups, and other entities from one domain to potentially access resources (files, directories, printers, plotters, and so forth) in another, provided, of course, that the appropriate access mechanisms (e.g., shares) and sufficient permissions are in place. Trust is an essential element in characterizing domains that are linked together to form trees or forests. These domains may be either in "Mixed Mode" or "Native Mode," as the next section explains.

Mixed Mode Versus Native Mode

Domains can be deployed in two modes: "Mixed Mode" and "Native Mode." In Mixed Mode, a domain contains both W2K and NT DCs, or has all W2K DCs, but nobody has migrated the domain to Native Mode. In Native Mode, a domain contains all W2K DCs and the domain has been migrated to this mode.

Native Mode is better from a security standpoint in that certain security-related functions (such as Kerberos authentication, a very strong type of network authentication, as explained shortly) are available only in this mode. The primary downside of Native Mode is that functionality is much more complex than in Mixed Mode. Complexity normally requires greater time and cost in planning and design; additionally, complexity generally makes security more difficult to achieve.

Domain Controllers

DCs are a special type of server used for controlling settings, policies, changes, and other critical facets of W2K domain functionality. In W2K mixed mode, DCs may consist of both W2K and NT servers. One W2K server must serve as a primary domain controller (PDC) in mixed mode, however. A PDC receives changes, such as changes to the authentication database, and replicates them to the other DCs within the domain. In W2K Native Mode, however, there is no PDC per se; all DCs are capable of picking up and replicating changes to the other domain controllers. Every DC in a Native Mode deployment holds a copy of Active Directory. In W2K Mixed Mode (or, in an NT domain), if the PDC crashes, some degree of disruption invariably occurs. In W2K, however, if any DC crashes, there is no particular problem—all DCs more or less function as equals to each other.

Active Directory is so important in understanding how W2K works that it merits further examination. The next section describes Active Directory functionality in greater detail.

Active Directory

Each object in a W2K tree or forest has an X.500 compliant distinguished name (DN), one that uniquely refers to the object in question (e.g., /O = Internet/DC = COM/DC = Example/CN = Users/CN = Jill Cooper). Each object also

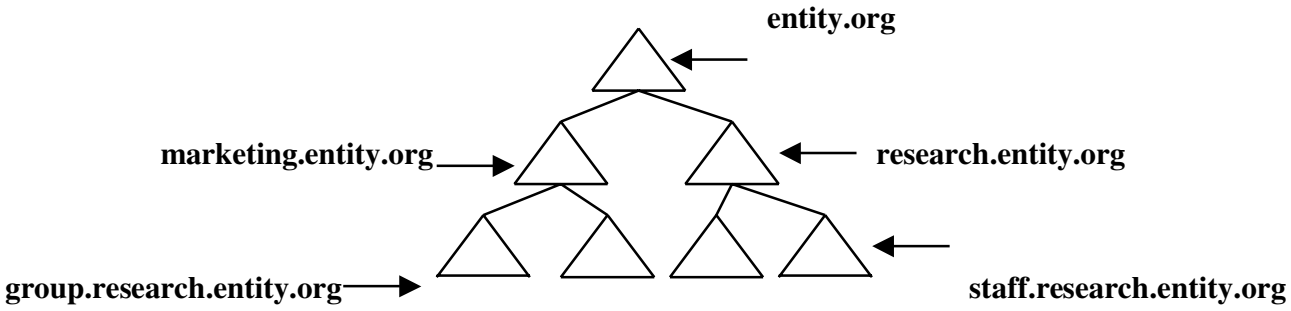


Figure 1: Example how name space is organized within a tree.

has a Globally Unique ID (GUID), a 128-bit identifier unique to the object within a particular namespace. X.500 properties and naming conventions are beyond the scope of this chapter; in brief, they provide an orderly way to organize and refer to objects. The X.500 directory structure is detailed and cumbersome; consequently, the Internet Engineering Task Force (IETF) created the Lightweight Directory Access Protocol (LDAP) to provide a kind of scaled-down, simplified version of X.500. W2K Active Directory is based on LDAP.

Active Directory objects are organized in various manners. Each object has one or more attributes. "Containers" are higher-level objects that hold other objects. Directories, for example, are containers that hold one type of object called "files" and another called "directories." The types of objects that that each container holds and the properties (e.g., names) of the objects are determined by the "schema."

Microsoft designed Active Directory with the goal of reducing barriers to locating and accessing resources throughout a tree or forest, regardless of whatever network boundaries (e.g., separation of networks from each other) exist. Each machine within a tree or forest contains objects (resources). The Global Catalog service enables users and programs that run on users' behalf to discover available resources within a tree or forest. When a trust link between domains is established, Global Catalog services extend across domain boundaries. When a request to access a resource occurs, the Domain Name Service (DNS) not only resolves hostnames into IP addresses and vice versa, but also resolves objects—that is, when given an object name, it identifies exactly which object it is. In effect, therefore, DNS is also the locator service for Active Directory. Service Resource Records (SRRs) are the basis for locating services and objects and to keep DNS tables up to date. Dynamic DNS (DDNS), a service available in more recent releases of BIND, updates Service Resource Records (SRRs) to ensure that Global Catalog and other directory service-related services can perform their functions properly. DDNS also registers systems with dynamic addresses that connect to the network.

Replication of Active Directory changes involves several steps. The update process (e.g., when a user changes a password or when an Administrator adds a new user to a group) begins when an update in the copy of Active Directory within the DC that receives a change occurs. Each DC that receives updates becomes part of a "replication topology" that specifies the particular connections within a tree or forest formed to synchronize the contents of Active Directory within each DC. At the designed time interval, the connections are established and updates are sent to each DC within the tree or forest. The Update Sequence Number (USN), a 64-bit number associated with an object, and Property Version Number (PVN), a version number for each object and the object's attributes, for the changed object are both incremented by the DC that records the update. Additionally, the DC captures the timestamp for the change.

Each DC updates its copy of Active Directory if the USN and PVN are higher than the values it has for the object in question. In case of a "conflict," that is, if there has been more than one change to the same object, the change with most recent timestamp will be recorded in any DC's copy of Active Directory. Each DC that initiates one or more updates and each DC that receives these updates constitutes a "replication partner." Several mechanisms are in place to protect against one or more rogue machines from replicating bogus changes—replication partners authenticate to each other, changes on every DC are tracked, and access control mechanisms ("permissions") determine who can modify Active Directory objects.

Group Policy Objects (GPOs)

GPOs are a collection of configuration settings related to computer configuration and user profiles. They provide a powerful way to store and flexibly apply policy settings. Several types of GPOs exist, including the following:

- Local GPOs (LGPOs), intended mainly for computers that are not part of any Active Directory domain
- Active Directory group policies, designed to be linked to various Active Directory containers, such as sites (which are explained shortly), domains, or OUs
- System Policies—System Policies are legacy groups of settings from NT System Policies if NT domains have been migrated to W2K systems.

Many different GPOs can be created and linked—some at the OU level, others at the domain level, others at sites (subnets or groups of subnets used in controlling replication of Active Directory changes as well as for network administration), and still others at the local level. GPOs are applied in a predictable order. For computers, any local computer GPOs are applied first, then site-linked computer GPOs, then computer GPOs linked to domains, and then local-linked computer OUs. The last GPO applied normally is the one with the settings that go into effect. This means that if there is a domain policy governing, say, account lockout parameters and a local policy governing the same, the domain policy settings will be the effective settings. The same basic principle applies to user GPOs—local GPOs that apply to users are applied first, followed by site-linked GPOs, and so forth. Another way of saying this is that OU-linked GPOs normally have precedence over all other levels, followed by domain-linked GPOs, followed by site-linked GPOs, followed by local GPOs (see Figure 2).

There is one important exception to the principle, however. In terms of place within the object hierarchy, sites are above domains, and domains are above OUs. Parent OUs are always above child OUs, too. If someone with sufficient privileges (e.g., a Domain Administrator) sets a "No Override" for a GPO that is linked at a higher level within this hierarchy, conflicting settings of GPOs set a lower levels will not apply (as shown in Table 1). The "No Override" setting thus becomes the good way for Domain Administrators to "gain the upper hand" by linking GPOs to domains and OUs, then setting a "No Override" on domain-linked GPOs. This, for the most part, ensures that any OU administrators will not be able to negate domain group policy.

If there are multiple GPOs within any single level of precedence (e.g., OU level), the policy that has been most recently linked to that level is by default the one that is

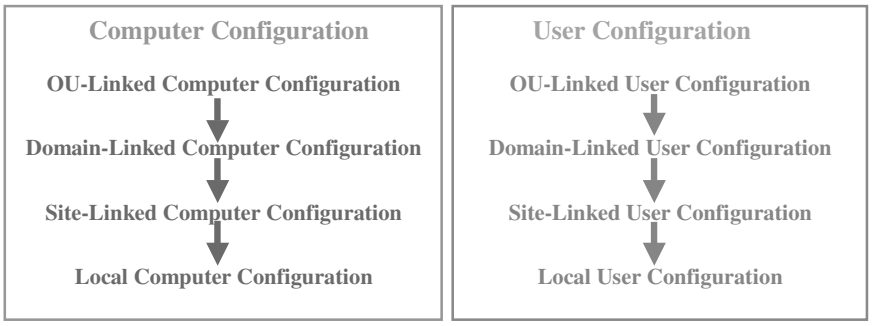


Figure 2: Precedence of GPOs at different levels.

applied. So the default GPO that is linked to a domain will be overridden by linking a new GPO to the same domain. Still, a Domain Administrator or someone else with sufficient privileges can reverse the order of precedence—the default GPO can go into effect simply by using the Group Policy Editor to reverse the order. Note that in Figure 3, there are two policies, EES Policy and Default Domain Policy, that are linked to the domain ees.test. EES Policy is listed first and will prevail over the Default Domain Policy link. However, by highlighting EES Policy in the Group Policy sheet shown in Figure 3 and then clicking "Down," the Default Domain Policy can be made to prevail.

Further complicating the situation is the fact that GPO settings can be inherited, for example, from one site container to its children or from one OU to its children. "Block Inheritance" settings can be in place at the level of children containers, however. "Block Inheritance" does exactly what it implies. But if there is a "No Override" at the higher level container (e.g., a parent site or OU), the "No Override" prevails; GPO settings from the parent are put into effect at the level of the child containers. Furthermore, inheritance does not work from one domain to its children. To have the same GPO settings for a parent domain and its children, therefore, it is necessary to link all the domains to the same GPO.

GPOs can profoundly affect W2K security. Consider password policy, for instance, as shown in Figure 4. Settings such as minimum password length and password complexity (i.e., whether passwords can consist of any set of characters or whether they must be constructed according to specific rules, e.g., that they may not contain the username and must include at least three of the following four categories of characters: uppercase English characters, lowercase English charact, numerals, and special characters such as "&" and "/") are embedded in GPOs. These settings affect how difficult to crack W2K passwords are. GPOs can be applied to a wide variety of entities, including accounts, local computers, groups, services, the W2K Registry, the W2K Event Log, objects within Active Directory, and more.

Table 1 How "No Override" Works

Level	"No Override" Applies to GPOs Linked to
Site	Child Sites
	Domains
	OUs
	Child-OUs
Domain	OUs
	Child-OUs
OU	Child-OUs

Accounts, Groups, and Organizational Units

As in NT, each W2K system has a default local Administrator account, the built-in superuser account for administering that system. A default Administrator account also exists within each domain for the purpose of administering systems and resources throughout the domain. Additionally, there is a default local Guest account and also a domain Guest account, both of which (fortunately) come disabled by default. Any additional accounts must be created by people or applications with the appropriate level of rights.

W2K groups are more complicated than accounts. W2K has four types of groups: local groups (for allowing access and rights on a local W2K system), domain local groups (which can encompass users or groups from any trusted domain), global groups (which can allow access to

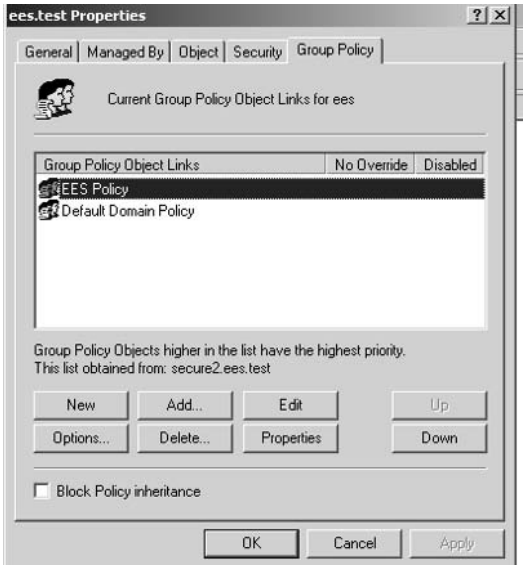


Figure 3: Viewing Group Policy Object Links for a domain.

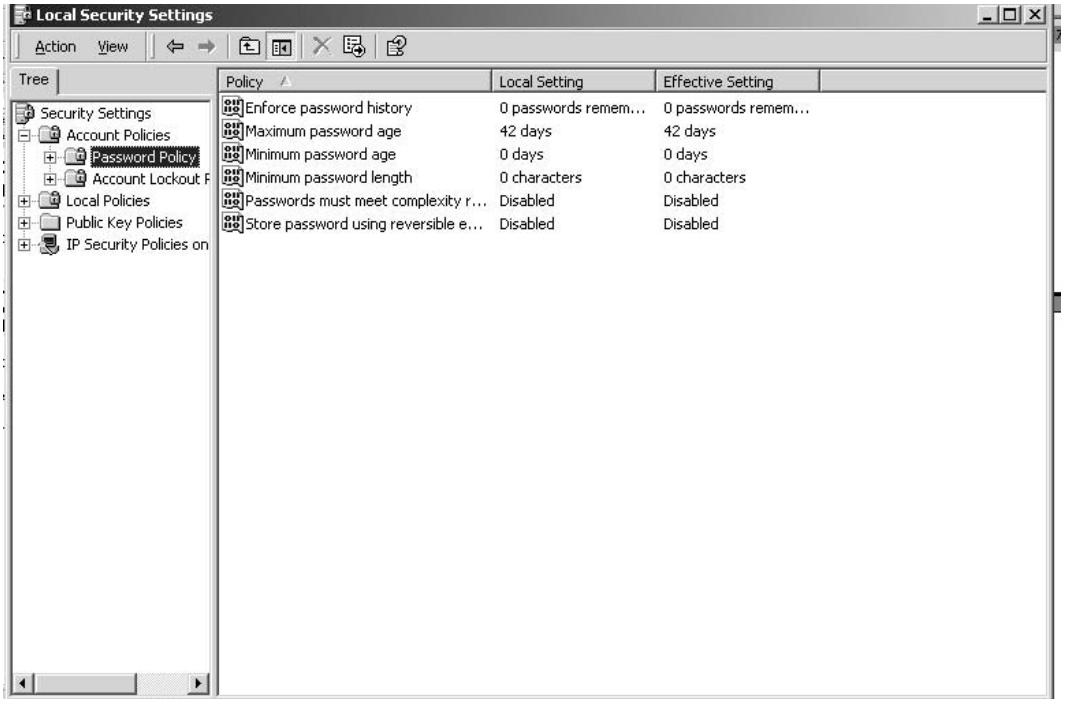


Figure 4: Password Policy settings.

resources in the domain or forest where they exist and are backward compatible with NT global groups), and (only in Native Mode) universal groups (which can consist of users and groups from any Native Mode domain within a tree or forest). Universal groups provide the most flexible way of forming groups and providing access to them at the risk of potentially allowing too much access to these groups unintentionally.

Some types of groups can be included within other groups. Group inclusion means that any users from one group can also become members of another group by adding the first group to the second. For example, global groups can be added to domain local groups and in a Native Mode domain, global groups can even be included in other global groups.

W2K's group inclusion properties provide a very convenient way of setting up access to resources, especially when trusted access is required. System administrators can, for example, include a universal group from another domain in a domain local group in their own domain to give users in the other domain the access they need. The users in the universal group from the other domain will have the same access permissions to the resources in question as the users in the domain local group. Table 2 lists

Table 2 Default Groups in W2K

Global Groups	Domain Local Groups in DCs	Local Groups in Workstations and Servers
Domain Administrators	Administrators (Local)	Administrators (Local)
Domain Users	Account Operators	Backup Operators
Domain Guests	Server Operators	Guests
Certificate (Cert) Publishers	Backup Operators	Power Users
Domain Computers	Print Operators	Replicator
Domain Controllers	Replicator	Users
Group Policy Creator Owners	Users	Interactive Users
Enterprise Controllers	Guests	Network Users
	Interactive Users	Everyone
	Network Users	Creator/Owner
	Everyone	Dial-up
	Creator/Owner	Batch
	Dial-up	Terminal Server Users
	Batch	
	Terminal Server Users	

the default domain and local groups in DCs and also in workstations and servers in W2K.

The Everyone group consists of all users on a given system, regardless of whether they have been authenticated. Fortunately, the potentially dangerous Everyone group is at least not afforded elevated privileges. Groups such as Interactive Users, Network Users, Dial-up, Batch, and Terminal Server Users are volatile groups. When users are engaged in certain tasks, they are included in these groups. When they are done with the tasks, they are removed. For example, someone who performs a local logon into a system is included in the Interactive Users Group as long as that user stays logged on locally. Additionally, some groups even apply to an entire tree or forest. For example, Enterprise Controllers consist of every DC in an Active Directory implementation.

Privileges

The previously mentioned default local Administrator account and default Domain Administrator account have superuser privileges—full privileges, meaning that while logged into this account someone can create or delete accounts and groups (unprivileged and privileged); disable accounts; add new users to groups; set the system time; make backups; take ownership of every file, folder and printer; create or delete shares to folders or devices such as printers; set up and run a scheduled job; unlock a locked computer; read and purge the Security Log, and many other things. Any account that is a member of the Administrators group on a local system has the same privileges as the default local Administrator account. The default domain Administrator account also has Administrator privileges, but they apply to every server and workstation within the domain in which this account exists. Anyone who is a member of Domain Administrators (of which the default Domain Administrator account is initially the only member, but others can be added) can use Administrator privileges on every machine within a domain.

W2K, like NT, has default groups that have some but not all Administrator privileges. Account Operators, for example, can create, disable, and delete any account that does not have elevated privileges as well as perform other tasks. Server Operators can perform many server administration tasks, including setting system time, logging on locally, and others. Backup Operators can backup systems as well as others. Print Operators can create and delete print shares, assign permissions for shares, install or delete print drivers, and also engage in a few other system administration tasks.

Organizational Units (OUs)

OUs are an important new feature of W2K Active Directory. OUs are in the most basic sense groups that are part of a hierarchical structure, with some groups at a higher level than others in the structure. The root OU is the uppermost one in this structure; OUs can exist at other levels of this structure as well. Any second-tier OUs, OUs immediately below the root OU, will all have the root OU as their parent OU. OUs are not unique to W2K, however; other network operating systems that adhere to X.500 or LDAP standards such as Novell Netware 4.X and up have OUs, for instance.

OUs can be used very advantageously. In W2K, any OU can be assigned conventional privileges or "rights" (also see the next section, which covers privileges) and/or "delegated rights," the capability to administer that OU by engaging in tasks such as adding users to the OU. Default children OUs inherit the policy settings of their parent. However, policy settings can be blocked for any OU, allowing different policies to be assigned to children than to their parent OU. Additionally, when it comes to delegated rights, a child OU can never have more delegated rights than its parent. These properties and features can help guard against rights proliferation in which too many users have too many privileges, which translates to a security catastrophe waiting to happen.

Access Permissions

NT featured version 4 of the NT File System (NTFS). W2K features version 5, or NTFS-5, which offers many more permissions than does NTFS-4, allowing precise control over levels of access to resources. There are 14 "base" or individual permissions and 5 combined permissions, each of which includes a number of base permissions. Each permission includes both an Allow and Deny setting. So, for example, one user could be allowed to Read Folder/Read Data in a certain folder, and another user could be assigned the deny setting for the identical permission for the same folder, preventing the second user from being able to read the folder and the data therein.

The FAT32 file system is also available, but this alternative file system has nothing to offer as far as security goes. There are, for example, no access permissions in FAT32. FAT32 features attributes such as read-only, but these attributes are easy for an everyday user to change. NTFS-5 also has some nice built-in reliability- and performance-related features.

Inheritance also applies to NTFS permissions and ownerships in W2K. Suppose that a subfolder or file is created below a parent folder. By default, a newly created child folder or file will inherit the permissions of the parent folder. It is also possible to block inheritance for any child folder or file. When an access request occurs, the Security Reference Monitor (SRM), an important subsystem within W2K, obtains information about the requesting user's security identifier (SID), groups to which the user belongs, and ownership of resources. The SRM next obtains the access control entries (ACEs) for the resource in question and evaluates them in a defined order. The SRM evaluates any Deny Non-inherited ACEs first, and then if there are no such ACEs evaluates any Allow Non-inherited ACEs. If there are no Allow Non-inherited ACEs for that resource, the SRM next evaluates any Deny Inherited ACEs, and if there are none, finally any Allow Inherited ACEs. If there is more than one ACE for one type of ACEs, e.g., Deny Non-inherited, the most recently created one is applied.

Kerberos

Kerberos provides strong network authentication both by authenticating users in a manner that keeps passwords

from going across the network and encrypting sessions and providing users with tickets ("service tickets") that enable users to connect to servers to access resources and services therein. Kerberos security is based on "Key Distribution Centers" (KDCs), servers that store user credentials and set up encrypted sessions on behalf of users who need to authenticate and then access resources and services. Each KDC distributes a unique, short-term session key for the client and KDC to use when they authenticate each other. The server's copy of the session key is encrypted in the server's long-term key. The client's copy of the session key is encrypted in the client's long-term key (which is usually based on the user's password).

In Kerberos, when a client wants to connect to a server (e.g., via a share), the following chain of events transpires:

- The client sends a request to the KDC.
- The KDC sends the client two copies of a session key. The client's copy of the session key has been encrypted using the key that the KDC and the client share. The server's copy of the session key and data concerning the client are contained in a "session ticket" that then becomes encrypted with the key that the KDC shares with the resource server that the client wants to access.
- Once the client receives a reply from the KDC, the KDC removes both the ticket and the client's session key and then caches them.
- When the client wants access to the server, it transmits a message containing the ticket and an authenticator that contains data about the user and a time stamp from the client to the resource server. The ticket will still be encrypted by the server's secret key; the authenticator will be encrypted by the session key.
- The KDC now sends a user ticket (often termed the "Ticket Granting Ticket"—a TGT) to the client.
- The client fetches the appropriate logon session key from its cache, using this key to create an authenticator and then sending the authenticator, the user ticket, and a request for a service session ticket to the KDC.
- The client sends both the authenticator and user ticket back to the KDC. The KDC responds by giving the client a server session key.
- The client needs to send the service ticket that is encrypted with the server session key to the resource server.

In W2K Native Mode each DC also functions as a KDC. Kerberos not only authenticates users and authorizes access to resources and services in Native Mode, but also serves as the basis for trust relationships between domains in W2K Native Mode. When trust is established between Native Mode domains, Kerberos keys for each domain are sent to the other domain for each KDC within it to use in authenticating and authorizing trusted access for users in the first domain. Another nice thing about Kerberos is that it is almost entirely transparent to users. Kerberos works only in Native Mode, however; this consideration provides strong impetus for migrating to this mode.

Security Support Provider Interface (SSPI)

SSPI consists of a Win32 interface between security-related "service providers" (dynamic link libraries or DLLs) and applications that run at the session level of networking as well as between other types of authentication packages. SSPI supports a variety of interfaces, enabling applications to call security providers to obtain authenticated connections. SSPI is potentially a considerable advantage for security in W2K systems because it provides an interface for third-party authentication products, such as the products developed by smart card vendors. Third-party authentication is much stronger than conventional, password-based authentication in that third-party authentication generally requires "something that you have" or "something that you are" plus "something that you know" (e.g., a Personal Identification Number or PIN) instead of only "something that you know" (i.e., a password).

Auditing

W2K can provide up to six types of logging, depending on the particular types that the system administrator enables. Types of logging include the following:

- System Logging—this reports events concerning errors, warnings, and information about the status of system operations and is nonconfigurable.
- Security Logging—this configurable log captures data about successful and failed access to objects such as files, directories, and printers, logons and logoffs, use of rights, policy changes, and so on.
- Application Logging—this log, which is configurable by application programmers, records application-related events (e.g., such as when Norton AntiVirus finds and eradicates a virus).
- Directory Service Logging—this configurable logging capability captures access (reads, writes, deletions, and so forth) to Active Directory objects.
- DNS Server Logging—various DNS-related events are recorded by the DNS Server logging capability, which is configurable.
- File Replication Logging—this configurable logging capability reports events related to Active Directory replication.

Of these six types of logs, the Security Log (as its name implies) is the most fundamental to security. The Security Log can be configured to capture successful and failed events from each of the following nine event categories:

- Audit account logon events
- Audit account management (e.g., creating, disabling and deleting accounts; group changes, and so on)
- Audit directory service access
- Audit logon events (e.g., every service logon)
- Audit object access
- Audit policy change
- Audit privilege use

- Audit process tracking (e.g., user attempts to start or stop programs)
- Audit system events (e.g., system startups and shutdowns)

GPOs can be used to set the audit policy for all the DCs within a domain as well as for member servers and workstations. Additionally, property settings for each of the types of logs determine the maximum size of each log and the retention method (e.g., whether to overwrite log entries only when the maximum log size is reached or not to overwrite events, that is, to clear the log manually).

Encrypting File System (EFS)

EFS provides encryption of folders and/or files stored on servers and workstations. EFS encryption is an advanced attribute for each folder and/or file. When a user enables encryption for a file, for example, a file encrypting key (FEK) is used to encrypt the file contents. When the user accesses the file (e.g., through an application), the FEK (which is used in connection with secret key encryption) decrypts the file. When the user finishes accessing the file, the FEK once again encrypts it. A key encrypting key (KEK), one of a public–private key pair, is used to encrypt a copy of the FEK. If something goes wrong, for example, if the KEK is deleted, authorized persons (by default, Administrators) can access the Data Recovery Agent snap-in, which uses the other key of the key pair to decrypt the FEK. Unfortunately, EFS in W2K is beset with a number of problems, including the necessity of sharing a user's FEK with others when more than one user needs to access an EFS-encrypted file. Despite the potential utility of folder and file encryption, the use of EFS in W2K thus is not generally advisable.

Encryption of Network Transmissions

W2K offers a number of ways to encrypt data sent over the network, including IPsec, the Point-to-Point Tunneling Protocol (PPTP), and other methods. IPsec is the secure IP protocol that features an authenticating header (AH) and encapsulated security payload (ESP). The AH provides a cryptographic checksum of the contents of each packet header that enables machines that receive "spoofed" packets, i.e., packets with falsified source addresses, to reject them. The ESP provides encryption of the data contents of packets, such that if anyone plants a sniffer on a network, the perpetrator cannot read the packet contents in cleartext. W2K provides IPsec support, although its implementation of the IPsec protocol limits the range of other systems with which W2K systems can set up IPsec sessions. W2K policy settings allow system administrators to set variables such as the conditions under which IPsec is used, the strength of encryption, and others. PPTP can also provide confidentiality of data sent over the network, although PPTP cannot verify the integrity of packets.

Routing and Remote Access Service (RRAS)

RRAS, another important W2K service, can be used to manage parameter settings for the W2K Remote Access Service (RAS), PPTP, and the Layer 2 Tunneling Protocol (L2TP). Among other things RRAS can be used to elevate security, in that this service can fix the method of authentication to be used (Kerberos, the older NTLM authentication method, and so forth) as well as filter and log incoming IP packets. IP packet filters can selectively determine whether packets will be received and/or forwarded on the basis of source IP address, destination IP address, and type of protocol. RRAS also allows system administrators to log all incoming IP traffic, something that is potentially very useful in identifying and investigating remote attacks.

Certificate Services

The Advanced Server version of W2K also offers certificate services. These include creation and release of X.509v3 certificates and even Public Key Infrastructure (PKI) capabilities. PKIs provide a hierarchical structure of Certification Authorities (CAs) that issue and validate certificates.

Distributed File System (DFS)

DFS is a function that enables system administrators to create and administer domain shares through a centralized function on each DC. DFS also allows administrators to assign permissions to shares, thus potentially limiting the level of access to resources throughout each domain.

Microsoft Management Console (MMC)

The MMC in W2K features "snap-ins," convenient objects that allow control of settings (group policy settings, in particular). Some of the snap-ins allow control of certificates, others are for computer management, others are for the event viewer, others are for group policy, and still others are for security templates. Security templates provide groups of settings that affect security and can be used to either evaluate the security level or to change unsafe settings to ones that are more suitable for security.

The services, functions, and properties discussed in this section are not the only ones that W2K offers, but they represent some of the most important services from a functionality and security standpoint. In the next section, I discuss the strengths and weaknesses of W2K security.

HOW SECURE IS W2K?

After all the problems that organizations and the user community have had with NT security, it is important to ask how secure W2K really is. The question really breaks down into two questions, however: (a) How secure is W2K out-of-the-box? and (b) How high a level of security can W2K achieve?

How Secure Is W2K by Default?

Some of same security-related problems that plagued NT are still present in W2K systems immediately after an installation of W2K. The permissions for the critical %systemroot%\ directory, for example, allow Full Control to Everyone in W2K. Additionally, in W2K Server and W2K Advanced Server, the IIS Admin Service runs

by default. Anyone whose system was infected by Code Red, Nimda, or another worm will appreciate the dangers of running a vulnerability-ridden Web server, namely, the Internet Information Server (IIS), by default. In other respects, W2K is more secure than NT was by default. When unprivileged access to Active Directory objects is necessary, for example, access is by default granted to Authenticated Users rather than the dangerous Everyone group. The point here is that W2K may be somewhat more secure than NT out of the box, but leaving W2K settings as they are is a huge mistake if security is a consideration. W2K systems need significant work to run securely.

Major Types of Vulnerabilities

W2K has had more than its share of security-related vulnerabilities. One of the most significant is a weakness in the way password representations for each account are created. The encryption process produces password representations that, by default, are relatively easy to crack using dictionary-based cracking techniques. If Service Pack (SP) 1 or higher is installed, the syskey utility is automatically installed. syskey adds a random, initial 128-bit encryption step before password representations are created, making password cracking more difficult.

Another significant set of vulnerabilities concerns susceptibility of W2K systems to denial of service (DoS) attacks. Programs for many W2K services are not particularly well written. An attacker may send input to these services that contains out-of-range parameters or that exceed memory limitations. The result is often DoS—either the programs or the W2K system itself—will crash. The W2K telnet server, for example, has a bug that will cause it to crash if certain types of input are sent to it. Similarly, massive Simple Network Management Protocol (SNMP) input may cause the receiving system to go into a buffer overflow condition in which there is too much input for available memory. This problem will normally cause a W2K system to crash, but if the excessive input is specially crafted, it is possible to execute rogue commands and programs on a system.

Some of the services that run in W2K systems pose much higher levels of risk than others. W2K Terminal Services, for example, provide a convenient way for users to connect remotely to other systems if these services are not properly configured, protected, and patched. The same is true for the W2K telnet server, the IIS Admin Service, SNMP, and many others.

Another vulnerability has been briefly mentioned earlier in this chapter. The default Administrator account is an attractive target for attackers in that by default it does not lock after an excessive number of unsuccessful logons. Additionally, this is a well-known account—one with a well-known name. Furthermore, being able to break into this account provides superuser-level privileges to attackers. Lamentably, many successful attacks on W2K are attacks in which perpetrators have broken into the Administrator account, which may not even have had a password.

Other major types of vulnerabilities in W2K that have been identified include the following:

- Bugs that result in privilege escalation; the consequences are particularly severe if the attained privilege level is Administrator
- Flaws that allow impersonation of other users (e.g., by allowing someone to run a process in another user's container)
- Ability to exceed allocated disk quotas (e.g., by repeatedly appending content with a certain number of bytes to files)
- NetBIOS-related vulnerabilities (in Mixed Mode; this layer of networking is beset with many security-related problems, including providing a wealth of information about systems, users, and current sessions to potential attackers)
- Bugs that can give attackers unauthorized access to Active Directory objects
- Poorly protected dial-in connections that require only a "normal" password for access instead of something stronger, such as smart card or biometric authentication (unauthorized dial-in access is one of the greater threats in any operating system, W2K included)

Microsoft has fixed most of the vulnerabilities mentioned here, as well as others that have been posted to newsgroups such as bugtraq. Microsoft generally initially produces a hot fix to repair a vulnerability or sometimes a set of vulnerabilities. Sometime later Microsoft releases either a set of bundled hot fixes or an SP that incorporates previously released hot fixes.

Now that we have explored the major types of vulnerabilities in W2K, let's turn to another, related consideration, namely, how secure W2K systems can be.

How Secure Can You Make W2K Systems?

W2K has numerous features that can substantially boost its security. As mentioned previously in this chapter, however, you'll have to make a number of changes to W2K if you want it to run securely. W2K has great security potential, but to achieve that potential requires considerable effort. The most important consideration is achieving a baseline level of security.

Baseline Security Measures

Establishing at least a baseline level of security is essential if W2K servers and workstations are going to withstand even the most basic kinds of attacks. Baseline security requires implementing the most fundamental steps in securing a system or application, not implementing a complete (more perfect) set of measures. The intention is to make a system "just secure enough." Implementing the following measures in W2K systems will produce a baseline level of security:

Install W2K From Trusted Media—A Vendor-Provided CD

Ensure that your system's hard drive consists of a minimum of two partitions, C: and D: Use C: as the installation drive; this partition will contain critical system directories and files. Do not set up user shares to this partition. In workstations and member servers use D: to hold other

files and folders; set up user shares to D: as needed. In domain controllers use D: to hold Active Directory files and folders; do not set up user shares to this partition. Set up the E: drive in domain controllers to hold user files and folders, and set up the user shares to this drive that are needed to allow users to access the resources they need to access.

Format Each Partition as an NTFS Partition

If any volume is FAT-formatted, enter the following:

```
convert <partition letter>:/fs:ntfs
```

For example, to make the d: partition into NTFS partition, enter

```
convert d:/fs:ntfs
```

then reboot the system.

Ensure that W2K Systems Are Part of a Domain

As mentioned earlier, workgroups provide few barriers to attackers. To check whether your system is part of a domain or workgroup, right click on My Computer to Properties, and then click on Network Identification.

If your W2K system has been upgraded from Windows NT 4.0, that is, it is not a native installation, use secedit to bring the default level of security to the level that is present in a native installation. secedit allows W2K security templates to be used in analyze and configure modes. In workstations and member servers, change your current directory to c:\%systemroot%\security\templates, then enter a command such as the following:

```
secedit/analyze/cfg securews.inf
/db%TEMP%\secedit.adb/verbose/log%TEMP%
  \scelog.txt
```

Install the Latest Service Pack (SP) and Hot Fixes

On W2K workstations and servers, SP3 is the most recent one. You can obtain this SP from http://www.microsoft.com/windows2000/downloads/servicepacks/sp3/

The following related steps should also be taken:

- Install the latest hot fixes, many of which fix the most recently discovered security-related vulnerabilities.
- Download post-SP3 hot fixes from http://www.microsoft.com/download/
- Download and run HfNetChk. This free Microsoft-provided tool enables system administrators to determine whether all W2K hot fixes have been installed. This tool works in connection with NT, W2K, SQL Server 7.0 and 2000, IIS, and IE 5.1. This tool is run from a command line. HfNetChk can be obtained from http://www.microsoft.com/downloads/details.aspx?displaylang=en&FamilyID=34935A76-0B20-4F91-AODE-BAAF969CED2B.

Lock Down Access to the System Drive (and, in the Case of Domain Controllers, the Drive on Which Active Directory Resides)

In general, do not assign anything more than Read-Execute permissions to Everyone, but always assign Full Control to Creator Owner and Administrators.

- Assign Authenticated Users Read-Execute access to c:\%systemroot% (which normally is c:\winnt or C:\w2ksrv) and c:\%systemroot%\system 32
- Assign Everyone Read-Execute access to the sysvol, sysvol\sysvol, and ntds folders (wherever they may reside in the file system). Remove all access (*but do not assign any Deny access*) to c:\%systemroot%\repair for the Everyone group

Avoid Sharing Folders Whenever Possible

Allow Creator Owner and Administrator to have Full Control over each share. Remove Everyone's access (*but do not assign any Deny access*), then assign Authenticated Users the Change level of share access. Change, which allows users to add files and subfolders, modify data, and delete files and subfolders, will not necessarily be the level of access Authenticated Users will get, however. If NTFS permissions for the files and folders that users can access via the share are more restrictive (e.g., they may allow only a Read and Execute), they will determine the actual level of user access to these resources.

To check or change permissions for domain shares or to delete shares, go from Administrative Tools to DFS to the DFS root. Open up the tree under DFS root until you get to the share you want to get to, then right click to Properties. Go to Administrative Tools, then either Computer Management and Local Users and Groups or to Domain Security Policy, then Active Directory Users and Groups (depending on the particular version of W2K):

- Rename the default Administrator account to an innocuous name, change the account description to "User account," enter a ridiculously long (up to 104 characters) and as difficult to guess a password as possible. Write the password down on the piece of paper that you keep in your personal possession (e.g., in your wallet or purse whenever you are at work). Never share this password with others and do not leave the slip of paper on which this password is written anywhere where others might see it. Use the default Administrator account, which in W2K does not lock after excessive bad logon attempts, only for emergency access.
- Create one additional account that is a member of the Administrators group for yourself and another for each person who needs to administer your system. Create an unprivileged account for each Administrator also. Use the unprivileged account when you are engaged in normal activities such as Web surfing, obtaining ftp access, and downloading mail. Use the superuser account only when you are involved in system administration duties.
- Create a new, unprivileged account named "Administrator." Ensure that this account is in only the Guest group. Look at your logs frequently to determine whether

people are trying to logon to this account—a decoy account designed to deflect genuine attacks against your system.
- Leave the Guest account disabled.
- Severely restrict the membership in the Enterprise Admins, Schema Admins, and Administrators groups, all of which have an incredible amount of power.

Go to Administrative Tools, then either Domain Security Policy or Local Security Policy (depending on the particular version, workstation or server, of W2K) and then to Security Settings:

- Go to Account Policies, then Password Policy to set the following parameter values:

Enforce password history	24
Maximum password age	90 days
Minimum password age	5 days
Minimum password length	8
Passwords must meet complexity requirements	Enabled
Store passwords using reversible encryption	Yes (unless your domain has WaX and Me clients)

- Go to Account Policies, then Account Lockout Policy to set the following parameters:

Account lockout duration	60 min
Account lockout threshold	5
Reset account lockout after	60 min

- Go to Domain Security Policy, then Active Directory Users and Groups or Local Security Policy, then Computer Management (again depending on the particular version of W2K you are running). Find the Users and Groups Container and double click on it. For each user account, set the following Account Options
- User must change Password at Next Logon: ensure this is clicked whenever a new account is created to help ensure privacy of user passwords (User Cannot Change Password—*do not* click on this)
- Password Never Expires—do not click on this except in the case of the default Administrator account and special accounts that have been installed for the sake of applications
- Account Is Disabled—be sure to confirm that the following accounts are disabled: Guest, accounts of employees who are no longer with your organization, accounts of employees who are on leave, and (unless your system is running an IIS web server) the IUSR_ and IWAM_ accounts. Disable these accounts by clicking on "Account Is Disabled" for each if they are not already marked with a red "X."

Set the following Security Options by going to Administrative Tools, then either Domain Security Policy or Local Security Policy (depending on the version of W2K each system runs), then to Security Settings, then to Local Policies, and finally to Security Options. Double click on the Security Options container. Double click on the option of your choice to either enable or disable it.

- Enable "Security restrictions for anonymous" to prevent anyone who connects to a W2K system via a null session from being able to enumerate shares and SIDs (Security Identifiers)
- Enable "Clear Virtual Memory Pagefile When System Shuts Down" to protect against an attacker gleaning sensitive information from pagefile.sys if the attacker is able to gain physical access to a system and boot from a Linux or other disk
- *Do not* choose "Shut Down the Computer when the Security Log is Full," "Recovery Console: Allow Automatic Administrative Logon," and "Allow Server Operators to Schedule Tasks."

Set a Baseline of Logging

Go to Administrative Tools, then either Domain Security Policy or Local Security Policy (depending on the version of W2K your system runs), then to Security Settings, then to Local Policies, then to Audit Policy. Double click on the Audit Policy container to view the audit options. To enable any type of auditing, double click on the name and in the sheet that will appear (under Audit these Attempts) click on both Success and Failure. At a minimum enable "Audit account logon events." If you need higher levels of auditing, enable additional types of auditing such as "Audit logon events," "Audit account management," "Audit policy change," and "Audit privilege use."

Set Logging Properties for the Security Log

Go to Administrative Tools, then Event Viewer. Click on Security and right click to Properties. Set Maximum Log size to at least 5000K and (under "When maximum log size is reached") click on "Overwrite as needed."

Check Your System's Logs Regularly (Daily, If Possible)

Doing this will help determine whether your system has been attacked or if someone has tampered with it.

Ensure That Only the Bare Minimum of Services Needed Are Running

Disable any unnecessary services by going to Administrative Tools, then Services. Highlight the name of each unnecessary service, double click, then under Service Status click on Stop and under Startup Type set this to Manual. The following are services that are usually *not* needed in W2K:

- Computer Brower
- FTP
- IIS Admin Service (this is needed for IIS Web and FTP servers)
- Indexing Service
- Messenger
- Print Spooler (unless a local printer is attached to the system)
- Remote Access Service
- SNMP

- Telnet
- Windows Installer Service
- Worldwide Web Publishing Service (this is needed for IIS Web servers)

Ensure That Rights Are Given Only as They Are Needed

Check User Rights by going to Administrative Tools, then either Domain Security Policy or Local Security Policy (depending on the version of W2K your system runs), then to Security Settings, then to Local Policies, and finally to User Rights Assignment. Double click on the User Rights Assignment container. To assign or revoke a right, double click on the right of your choice, then add or remove the right to/from the user or group of your choice. Ensure at a minimum that the Everyone group *does not* have any of the following rights:

- Act as part of the operating system
- Add workstations to domain
- Back up files and directories
- Create a pagefile
- Create a token object
- Debug programs
- Enable computer and user accounts to be trusted for delegation
- Force shutdown from a remote system
- Increase quotas
- Increase scheduling priority
- Load and unload device drivers
- Lock pages in memory
- Logon as a batch job
- Logon as a service
- Log on locally
- Manage auditing and security log
- Modify firmware environment variables
- Replace a process level token
- Restore files and directories
- Shut down the system
- Take ownership of files and other objects

Install and Run Anti-Virus Software

A Final Caveat

It is important to not only *establish,* but also have to *maintain* suitable levels of W2K security. Security is an ongoing process. You cannot simply set certain parameters in a W2K or any other type of system and then forget about security. Good security requires inspecting systems to ensure that there are no unexpected changes in permissions, rights, directories, and files within directories. Anti-virus software has to be updated constantly if it is to be effective. Good security requires systematic monitoring of logs to spot and investigate suspicious activity. Good security also requires making full and incremental backups as well as an Emergency Repair Disk at appropriate time intervals. In short, good security for W2K or any other operating system is an ongoing process.

CONCLUSION

This chapter has provided the foundation for understanding W2K security capabilities, limitations, and solutions. W2K is a complex operating system. Its potential for security is higher than its predecessor, NT, yet its out-of-the-box configuration leaves much to be desired. This chapter cannot be considered complete coverage of the topic of W2K security. Entire books on the topic have been written (see Further Reading), yet even these do not cover everything pertinent to the complicated subject of W2K security. Some that are likely to be helpful in gaining a deeper understanding of W2K security include books by Bragg (BRAG00), Cox and Sheldon (COX00, Norberg (NORB00), Schultz (SCHU00), and Scambray and McClure (SCAM01). The recommendations in this chapter are designed to provide a baseline level of security in W2K. Recommendations for achieving higher levels of security are provided in other, longer documents (see http://www.cisecurity.org) and books such as the ones listed below.

GLOSSARY

Active directory A directory service that provides an infrastructure and related services that enable users and applications to locate and access objects and services throughout the network.

Containers Higher level objects that hold other objects.

Delegation Giving organizational unit–related rights to organizational units.

Distributed File System (DFS) A function that enables system administrators to create and administer domain shares through a centralized function on each domain controller and also allows administrators to assign permissions to shares.

Domain Name Service (DNS) A service that resolves IP addresses to hostnames and vice versa.

Domain A group of servers and (normally) workstations that are part of one unit of management.

Domain Controllers (DCs) Machines that hold information related to policies, authentication, and other variables.

Domain local groups Groups that can encompass users or groups from any trusted domain.

Encrypting File System (EFS) A system that provides encryption of folders and files stored on W2K servers and workstations.

Forests (As opposed to "trees") Trust-linked domains that are characterized by noncontiguous name spaces.

Global catalog A service that enables users and programs that run on users' behalf to discover available resources within a tree or forest.

Global groups Groups that can allow access to resources in the domain or forest where they exist.

Group Policy Objects (GPOs) A collection of configuration settings related to computer configuration and user profiles.

HfNetChk A free tool from Microsoft that enables system administrators to determine whether W2K hot fixes have been installed.

Hot fix Microsoft's term for a patch that fixes security and other types of problems in Microsoft products such as W2K.

Inheritance The default propagation of access rights and user rights (privileges) from higher level objects to child objects.

IntelliMirror A set of features that enable user data, applications, and computing environments to be available and user-specific settings to be applied.

IPsec A secure Internet Protocol that has an authenticating header and encapsulated security payload.

Kerberos A protocol that provides strong network authentication.

Key Distribution Centers (KDCs) Kerberos servers that store user credentials and set up encrypted sessions on behalf of users who need to authenticate and then access resources and services.

Lightweight Directory Access Protocol (LDAP) A protocol that provides a scaled-down, simplified version of X.500 directory services.

Microsoft management console A management tool that features "snap-ins," convenient objects that allow control of settings (group policy settings, in particular).

Mixed mode A deployment mode in which a domain contains both W2K and NT domain controllers, or has all W2K domain controllers, but nobody has migrated the domain to Native Mode.

Native mode A deployment mode in which a domain contains all W2K domain controllers and the domain has been migrated to this mode through an Active Directory setting.

NT File System (NTFS) A system conducive to strong access control—W2K offers version 5 of NTFS or NTFS-5.

Organizational Unit (OU) A "nested group," one that is either above or below other OUs (or both) in a hierarchy of OUs, with special properties that allow for delegation and inheritance of rights.

Primary domain controller A domain controller that receives changes, such as changes to the authentication database, and replicates them to the other domain controllers within the domain.

Replication The distribution of changes in Active Directory objects, properties, settings, and so forth from one domain controller to the others.

Schema An Active Directory characteristic that determines the types of objects that that each container holds and the properties (e.g., names) of the objects.

Security Support Provider Interface (SSPI) A Win32 interface between security-related "service providers" (dynamic link libraries, or DLLs) and applications that run at the session level of networking, as well as between other types of authentication packages.

Service Pack (SP) A set of bundled hot fixes.

Service Resource Records (SRRs) The basis for locating services and objects and to keep DNS tables up to date.

Syskey Microsoft's attempt to make it more difficult to crack passwords by adding an extra 128-bit encryption step in which passwords are encrypted before they are stored.

Tree A group of trust-related domains that form a contiguous name space.

Trust A property that potentially allows users, groups, and other entities from one domain to access resources.

Universal groups In Native Mode, groups that can consist of users and groups from any Native Mode domain within a tree or forest.

Workgroup A set of Windows and possibly other systems that are known to each and that facilitate access to each others' resources.

CROSS REFERENCES

See *Encryption; Internet Architecture; Internet Security Standards.*

FURTHER READING

Bragg, R. (2000). *Windows 2000 security.* Indianapolis, IN: New Riders.

The Center for Internet Security Web site. http://www.cisecurity.org

Cox, P., & Sheldon, T. (2000). *The Windows 2000 security handbook.* Berkeley, CA: Osborne.

Norberg, S. (2000). *Securing Windows NT/2000 servers for the Internet.* Sabastopol, CA: O'Reilly.

Schultz, E. E. (2000). *Windows NT/2000 network security.* Indianapolis, IN: New Riders.

Scambray, J., & McClure, S. (2001). *Hacking Windows 2000 exposed.* Berkeley, CA: Osborne.

Wireless Application Protocol (WAP)

Lillian N. Cassel, *Villanova University*

THE WIRELESS APPLICATION PROTOCOL (WAP)

WAP is a collection of protocol standards whose purpose is to enable communication between handheld wireless devices and Internet-based service providers. These service providers include Web servers providing content formatted especially for small wireless devices and other service providers who target wireless devices exclusively.

This set of standards builds on existing standards, reusing or modifying them where necessary to address the special needs of the handheld wireless community. The particular limitations of the wireless world include greatly restricted bandwidth, nonrobust connections, signal security limitations, the small screens on most devices, limited battery life, and restricted input options. Processor power and memory are other limitations for most target devices (Wireless Application Protocol Wireless Session Protocal Specification, 1999).

The Current Situation

The mobile telephone industry is currently suffering from disparities in standards that limit the usefulness of phones as they travel from one location to another. Currently, wireless system providers in the United States adhere to one of four different standards: IS-95 CDMA, TDMA, GSM, or IDEN. CDMA is code division multiple access and is the basis of the third generation (3G) wireless transmission technologies WCDMA and CDMA2000. TDMA is time division multiple access, in which each conversation has access to the carrier frequency only part of the time. There are a number of implementations of TDMA, including the digital American mobile system (D-AMPS), the global system for mobile communications (GSM), personal digital cellular (PDC), and the integrated digital enhanced network, (iDEN). GSM is uniformly used throughout Europe allowing systems to roam and retain service over large areas. GSM is also the standard in use in most African nations (Digital Mobile Phone Networks in Africa). PDC is the standard for cell phone access in Japan, where its widespread use makes it the second most used standard in the world. IDEN is a proprietary standard from Motorola. The lack of interoperability among the various standards is a serious limitation, restricting access as users move about (Dornan, n.d.a.).

An important characteristic of the communication mode is the basic model for connectivity. Connection can be either packet-switched or circuit-switched. Circuit-switched means that a connected user consumes bandwidth for as long as the connection is maintained. Most systems charge by connection time, reflecting the consumption of bandwidth. Packet-switched connectivity means that the device consumes resources only when actually sending or receiving a transmission. There is no reason to disconnect the device from the network. This always-on feature has significant consequences for wireless service providers. As a result, the most actively used wireless services are those in Japan, where cell phone use is pervasive and services are always available, without the need to establish a connection.

Mobile phone networks are moving toward advanced technologies, which will bring packet-based communication to the systems used outside Japan. Expectations are that the always-on characteristic will drive interest in and use of emerging applications available to mobile phone users. If the experience in Japan translates into usage patterns in the rest of the world, the demand for applications will grow at a great rate in the near future. Developed to run over any and all of the mobile device transmission methods, the wireless application protocol suite is the common set of standards that allows these applications to be available to all wireless devices.

MOBILE ACCESS TO THE WORLD WIDE WEB AND OTHER DATA RESOURCES

The World Wide Web has become a utility, taken for granted as an information and communication resource. Similarly, the mobile communication device—whether phone or personal digital assistant—has become a part of the environment for many people. It is natural to look to combine these two. Access to existing Web pages is one goal of emerging standards and technologies for handheld, mobile, wireless devices. However, these devices offer opportunities to provide services and information feeds beyond what the Web provides. Web pages are dynamic, but they are not very responsive to the location

of the visitor. Although there are location-specific services, such as yellow pages, most Web-based information is relatively static. A mobile user has different needs than a stationary browser. A mobile user might want to know what the traffic is just ahead or how to get from here to there. By combining the other features of the mobile device with a GPS system (or mobile positioning system "MPS" technology), an application could address those questions. When a mobile user asks for information about nearby restaurants, the response needs to be organized for effective viewing on the small screen and must accommodate the limitations of imperfect connections. When a mobile user queries his or her bank balance before making a purchase, strong encryption must protect the data transferred without unduly impacting the amount of time required to get the answer.

One class of applications geared specifically to mobile users is localized data push. Assuming the user gives permission for such intrusion, a store may broadcast notice of a sale or even a special offer to a mobile device user who happens to be driving or walking nearby. Thus, the information obtained will be offered by a company or organization to attract the attention of a potential customer and entice him or her with coupons or other offers.

Designing applications for the mobile device user requires criteria different from those for designing Web-based applications that will be accessed through large-screen devices with high-speed, reliable connections. Income for the service provider is likely to be in the form of a charge for the amount of information delivered rather than advertising revenue. That means that the information presented to the user must be efficiently designed and delivered. The small screen size of most mobile devices requires consideration in designing the information presentation. Having to scroll both horizontally and vertically to see the content of a message is difficult with devices typically held and controlled by one hand. User responses must be obtained without the use of a full keyboard and often without an effective pointing device. The design challenges are significant.

Standards

WAP must deal with two sets of constraints: limitations of wireless data networks and limitations of the handheld devices used to send and receive data. When compared to landline networks, wireless data networks, regardless of the technology used, are characterized by

less bandwidth,
more latency,
less connection stability, and
inconsistent availability.

Compared to desktop or even full-featured portable computers, handheld wireless devices are characterized by

small screens,
limited input options,
less memory, and
less powerful CPUs.

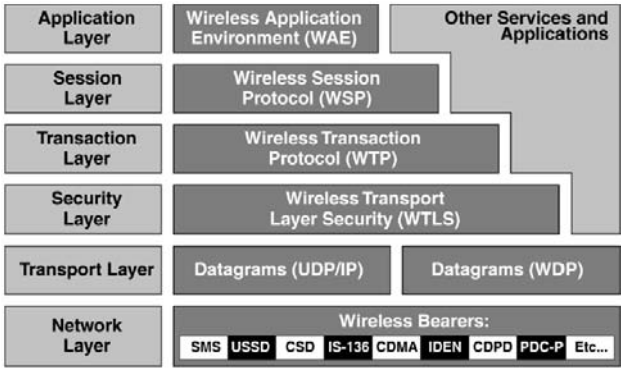


Figure 1: The WAP protocol stack (Wireless Application Protocol Forum, Ltd., 2002). © 2002 Wireless Application Protocol Forum, Ltd. Reprinted with permission.

Although technology will improve the situation, some characteristics will remain. Screen sizes will be small because an important feature of these devices is their small size and light weight. Battery life restrictions will limit CPU and memory size. WAP is a set of standards designed with these restrictions in mind. The WAP protocol stack is similar to the ISO OSI Reference Model (ISO/IEC, 1994) for the upper layers. Layers below the network layer are implied in the bearer protocols. Figure 1 shows the protocol stack with the WAP standards in place.

Wireless Application Environment

WAE is a framework for the development of applications that can be accessed from many types of wireless devices from a variety of manufacturers (Wireless Application Environment Specification, 2002). The goal is a structure in which operators and service providers can build their products with confidence of interoperability with a wide variety of devices and applications. WAE includes XHTML mobile profile [HTMLMP], WML, WCSS, WMLScript, WBXML, vCARD, and vCalendar. Each of these is summarized briefly below:

XHTML mobile profile W3C is migrating HTML to XHTML, making it an XML application, and at the same time making it modular. Applications can be built using just the modules that are appropriate for the target devices. Starting with module XHTML Basic, XHTML mobile profile adds extensions suitable for the mobile devices.

WML The wireless markup language, an XML-based markup language designed for use on devices characterized by low-bandwidth connections, limited memory and restricted CPU capacity, small screen, and limited user input interfaces.

WCSS The wireless cascading style sheet. Cascading style sheets are used in Web development to control display without sacrificing device independence. WCSS is specified for the features of small mobile devices.

WMLScript A scripting language, roughly similar to JavaScript, designed to run on small mobile devices.

WBXML A compact binary representation of XML documents intended to reduce the size of the files for transmission without losing semantic information.

vCARD and **vCalendar** Industry standards for sharing address and calendar information.

Wireless Session Protocol

WSP offers the application layer consistent interfaces for session services. Two modes are offered. A connection mode runs over the wireless transaction protocol and a connectionless service runs directly over a datagram transport service. Initial versions of WSP offer browsing capabilities including the equivalent of HTTP with implementation more suitable for wireless devices, plus facilities for long-lived sessions, session suspend/resume, data push, and capability negotiation (Wireless Application Protocol Wireless Session Protocol Specification, 1999; Wireless Transport Layer Security V, 2001).

Wireless Transaction Protocol

WTP provides services to allow browsing; specifically, WTP provides request response interaction between a client and a server. Unlike regular Web surfing with HTTP, WTP runs over a datagram service at the network layer. WTP enhances the service provided by unreliable datagrams in order to provide appropriate service levels to the higher layers, relieving the wireless device of the need to support all the services of TCP but providing a reliable service to the application. By using a datagram transport service, WTP eliminates the need for connection establishment and connection release activities. The protocol transmits a message as its basic unit, rather than a stream of bytes. WTP defines three classes of service:

Class 0 Unreliable invoke message with no result message,

Class 1 Reliable invoke message with no result message, and

Class 2 Reliable invoke message with exactly one reliable result message.

Class 2 is the basic invoke/response service. Class 0 is available for an invoke message that will not be retransmitted if delivery fails and for which no response is expected. Class 1 is used for an invoke message that must be delivered or re-sent. The variations allow an application to function with a minimum of data transmission required.

Wireless Transport Layer Security

WTLS exists to provide privacy and security between communicating entities in a wireless environment. Security includes both data integrity and authentication services. The services of WTLS are optional and their use depends on the nature of the application and the vulnerability of the data it transmits.

Datagrams

Datagrams are transport protocol mechanisms that offer best-effort delivery service without the high overhead required to provide highly dependable transport. Datagram transport protocols are lightweight, fast, and suitable for situations where the underlying network services are highly reliable or where the application will provide reliability and so services at the transport layer are redundant.

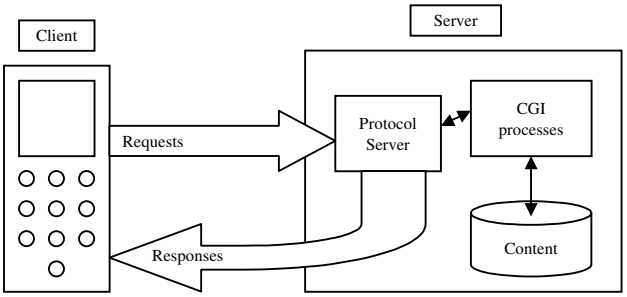


Figure 2: Direct interaction between the wireless client and the wireless application server.

Access to wireless applications can be provided by direct connection to a WAP server that is also a content provider, as illustrated in Figure 2. However, because most of the Internet content is present in Web sites designed to respond to HTTP requests and formatted in HTML that is often not suitable for wireless devices, standard access includes a gateway. The gateway receives requests using the WAP protocols and passes those requests to conventional Web servers using HTTP. The gateway reformats responses for display on the limited device and encodes the response for transmission using the WAP protocol stack. Figure 3 shows wireless client interaction with a conventional Web server.

WIRELESS MARKUP LANGUAGE

WML Basics

As we have seen, normal HTML-specified Web pages are seldom suitable for display on handheld wireless devices. The WAP protocol suite includes an XML language designed to specify items for display on these devices. The wireless markup language (WML) borrows heavily from HTML, but eliminates features not suitable for these devices and adds functions that address the screen size and user input options (Evans & Ashworth, 2001; Wireless Markup Language Version 2.0, 2001).

The first difference to notice is the paradigm of presentation. Where the Web metaphor is a page, the WML metaphor is a card. A single application display is a deck

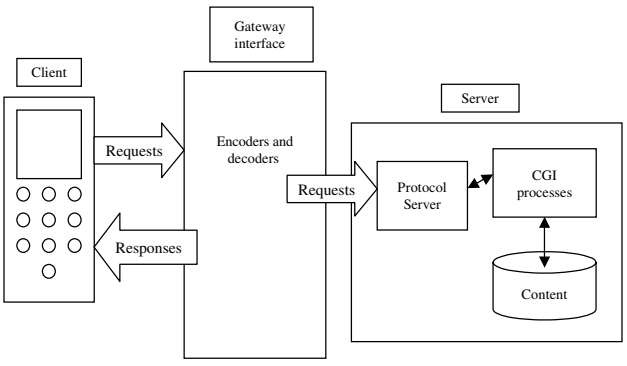


Figure 3: Wireless application client interaction with a conventional Web server.

```
<?xml version="1.0"?>
<!DOCTYPE wml PUBLIC "-//PHONE.COM//DTD WML
1.3//EN" "http://www.phone.com/dtd/wml13.dtd">
<!-- WML file created by Openwave SDK -->
<wml>
  <card id ="Start">
    <p align="center"> <b> City
information</b> </p>
   </card>
</wml>
```

Figure 4: First WML card specification.

of cards, much as people sometimes use in gathering research data in a library. Each card contains a small amount of content and may contain links to other cards in the same deck, to cards in other decks, or even to conventional Web pages.

An example application is the best way to illustrate and introduce WML. The example is developed using the Openwave SDK, available free from http://www.openwave.com. The simulated displays for a mobile phone are part of the SDK. The appearance of output from a WML specification will vary by the nature of the device on which it is displayed.

The example allows a client to obtain information about a particular city. The first screen identifies the application. The WML code for this initial card is shown in Figure 4.

The first two lines of this WML file reflect the XML connection. This identifies the XML version and allows valid compilation of the file. The actual WML specification begins with `<wml>` and ends with `</wml>`. Like other XML applications, each WML element has both an open and a close component. Notice that the `<p>` (paragraph) tag must be closed by `</p>`, unlike ordinary HTML. The `<wml> ...</wml>` tags mark the beginning and end of this deck. The deck may contain as many cards as the application requires. Each card is specified by `<card> ...</card>`. The `<card>` tag shows an ID for this card that can be used to reference it from other cards. The title of this card has been centered by an option in the `<p>` tag and made bold by the usual `<b> ...</b>`. Figure 5 shows the simulated output for this first WML example.

Figure 5: Simulated display of first WML code.

Next, the example adds options to the display. Each option will consist of a link to another card with further information on the chosen topic. This requires a list. The list in this application will allow the user to choose weather information, restaurants, museums, or information about public transportation. Each option in the list specifies the next card to visit if that option is selected. The entire select list goes inside a `<p> ...</p>` tag element. The code follows:

```
<select name = "listname">
  <option event = "link"> Label </option>
</select>
```

The syntax for the list in which each option links to another card in the current deck is

```
<?xml version = "1.0"?>
<!DOCTYPE wml PUBLIC "-//PHONE.COM//DTD
  WML 1.3//EN"
"http://www.phone.com/dtd/wml13.dtd">
<!-- WML file created by Openwave SDK -->
<wml>
  <card id = "Start">
    <p align = "center"> <b> City information
      </b>
</p>
    <p align = "center">
      <select name = "categories">
        <option onpick = "#weather">
          Weather Forecast </option>
        <option onpick = "#restaurants">
          Restaurant List </option>
        <option onpick = "#museums"> Museums
          </option>
        <option onpick = "#trans">
          Public Transportation </option>
      </select>
</card>
</wml>
```

The `listname` joins this set of options with a common title. `Label` is displayed on the user screen. `event` is the occurrence that causes this item to be selected from among the options available. `link` identifies the card to be displayed next. The following code includes a list of options for city information. Each link is relative; in fact it is a reference to a card in the same deck. The syntax for

Figure 6: First list attempt.

that reference repeats the HTML syntax for a jump to a location in the same page. Figure 6 displays the result.

Figure 6 shows the simulated output for this first list. Immediately, the constraints of the small screen become apparent. The wordiness of the descriptions is too much for this device to display well. In Figure 7 the descriptive text has is shorter and the list fits comfortably on the screen.

The user selects a list option by means of the arrow keys or by pressing the number corresponding to the chosen list entry. In Figure 7, the Museums entry is chosen. Pressing the button under the check mark will make the selection. To continue the application development, the additional cards are added to the deck. The empty cards are shown below:

```
<?xml version = "1.0"?>
<!DOCTYPE wml PUBLIC "-//PHONE.COM//DTD
  WML 1.3//EN" "http://www.phone.com/dtd/
  wml13.dtd">
<!-- WML file created by Openwave SDK -->
<wml>
  <card id = "Start">
    <p align = "center"> <b>
      City information</b>
</p>
    <p align = "center">
        <select name = "categories">
            <option onpick = "#weather">
              Weather
</option>
            <option onpick = "#restaurants">
              Restaurant List
</option>
            <option onpick = "#museums">
              Museums
</option>
            <option onpick = "#trans">
              Public Trans.
</option>
        </select>
    </p>
  </card>
  <card id = "weather">
  <card>
  <card id = "restaurants">
  </card>
  <card id = "museums">
  </card>
  <card id = "trans">
  </card>
  </wml>
```

The initial display looks no different, but now there is a card corresponding to each choice. The cards are currently empty, ready to be filled in with relevant information for this application.

Figure 7: Revised list to fit the screen.

WML Tasks

Forward and Backward Navigation

A WAP application keeps track of the sequence of cards displayed and also of how each card was reached. As a user views the cards in some order by selecting from lists or choosing to progress to the next card, the application records the identities of visited cards on a stack. Going back to the previous card consists of popping the stack. When the user goes back, there is no option to return to the card most recently viewed. In other words, the history mechanism is strictly a stack. Once popped from the stack, the historical reference is gone. The WML history structure stores only the identity and reference information for each card viewed; it does not store the content of the card. Thus, if a variable displayed on a card has changed value since the card was seen, revisiting the card will show the variable with its current value. On the other hand, if a card was first obtained by an HTTP post method the card will be refetched using the same values as in the original fetch. In summary, local variable changes will be reflected when old cards are redisplayed; if a card is obtained by an HTTP POST, revisiting the card will cause the original fetch operation to be repeated, using the same values.

Four tasks define the WML navigational behavior: `go`, `prev`, `noop`, `refresh`. Of these, `go` and `prev` provide forward and backward navigation. The `go` task causes a referenced card to be loaded and displayed. The `prev` task causes the history stack to be popped. The top card on the history stack is identified and displayed. (Results if there is no history are not defined in the standard.)

Do Nothing and Update Context

The `noop` task causes no processing. The `refresh` task causes the current card to be redisplayed using the current variable state and any processing under way stops. If the current card contains a timer, the timer is started on the `refresh` task.

If any task fails to complete, the current card remains in view. No changes are made to the browser context and no intrinsic event bindings are executed.

WML Events

Two types of events are included in the WML specification: intrinsic events and extrinsic events. Intrinsic events are internally generated; extrinsic events originate with an external source. The WML specification does not define any extrinsic events. Four intrinsic events are defined: `timer`, `enterforward`, `enterbackward`, and `pick`.

The `timer` event occurs when a timer goes off. A timer element measures elapsed time in tenths of a second. (A given device may be more or less precise in measuring tenths of seconds. However, the application is to consider a timer tick to represent a tenth of a second.) A card may have a specification of a timer and a corresponding event to occur after the timer expires. A timer begins to tick when the card is entered and continues to count down until it reaches zero. When the timer becomes 0, the specified event occurs.

For example, the following code will return the display to the first card after five seconds:

```
<onevent type = "ontimer" >
    <go href = "#Start"/>
  </onevent>
<timer = value = "50" >
```

`Enterforward` and `enterbackward` keep track of the way a card was reached. `Enterforward` occurs when the user reaches a card through any action except those that pop the history stack. `Enterbackward` corresponds to reaching a card by an action that pops the history stack. In this respect, WML provides state awareness that is missing in HTML. A WML application can tell if a card is reached by going back to it or if it is entered by a forward-moving choice.

An event binding maps a task specification to an event occurrence. For example,

```
<card id = "Third"
  onenterbackward = "First">
  <p> This is card Three </p>
</card>
```

will display "This is card Three if it is entered by choice or by forward sequence, but will jump to the first card (or the one labeled "First" to be exact) if it is reached by using going back."

`Event pick` appeared in the introductory example. There, `onpick` was used in the options nested within a select element. For example,

```
<option onpick = "#weather"> Weather
  </option>
```

mapped transfer to the card labeled "weather" when the user highlights this option and pushes the accept button.

In summary, intrinsic events occur in three categories. `Enterforward` and `enterbackward` map to reaching the card either by forward or backward navigation. Timer events map to a timer going off. `Pick` maps to an explicit user selection.

Triggering Events with Access Keys

One user activity is listed with the intrinsic events: selecting a choice from a list and pressing a key or button associated with acceptance. There are a number of buttons on the handheld devices used for wireless access. Events can be associated with others besides the `accept` key and there is another way to associate an event with pressing the accept key. (On the simulator used in examples, the `accept` key is below the screen and to the left. It is a physical button, not the onscreen simulated button. The onscreen button serves as a holder for a label for the real button, which the user must push.) The `do` element associates an action with pressing a specified button on the user device. For example,

```
<do type = "accept" label="Restart">
  <go href="#Start"/>
</do>
```

produces the display shown in Figure 8. The label on the accept button has been changed to say Restart. Pressing

Figure 8: Navigation button.

that button displays the first card in the deck, the one labeled Start. This device also has a button permanently labeled Bck, which corresponds to the Prev event in WML. Pressing that button backs up the display to the previous card, popping the history stack. The button to the right, just below the screen corresponds to the reset event in WML, and the options event. The Clr button corresponds to the delete event. Each of these buttons can be controlled by the WML code by substituting the correct event name for `accept` in the code example above. Substituting "reset" for "accept" in the code above will move the Restart label to the right button and pressing the right button will take the display back to the first card in the deck.

Variables

WML specifications describe content and navigation, not processing. However, there are occasions when the ability to store a value in a variable allows flexibility needed in the application. Variable names are case-sensitive, begin with a letter or underscore, and continue with 0 or more letters, numbers, or underscores. The variable reference syntax is specified using the extended Backus–Naur form used in XML as follows:

```
varref :: = ("$" varname) | ("$(" varname
  (conv)? ")")
varname :: = ("_" | alpha) ("_" | alpha |
  digit)*
conv :: = ":" ("escape" | "noesc" | "unesc")
alpha :: = [a-zA-Z]
digit :: = [0--9]
```

Parentheses may be used for clarity and are required where it is not otherwise clear where the variable name ends. Both `$(var)` and `$var` are legal references to the variable `var`. Parentheses would be needed for clarity in the following case: "Move next to the `$(var)`th card"

Because `$` is used as the referencing character for value substitution, any use of the dollar sign character must be escaped. This is accomplished by using `$$` wherever a real dollar sign is wanted. The need to use `$$` applies inside quoted strings as well as in other uses.

Variables are defined and given value by the setvar element of WML. For example,

`<setvar name = "idcode" value = "123">` creates a value called idcode and assigns it the value `123`. Variables receive a type as determined by the kind of value assigned.

Images and Tables

WML accesses images in the way familiar from HTML. Of course, the small display area limits the size and complexity of suitable images. WAP has a wireless bitmap graphic format (WBMP). Some devices will also accept images in other formats. The syntax for specifying an image for a WAP card is

```
<img src = "url" alt = "text" />
```

where url is the location of the image to be included and text is to be displayed if the device is unable to display the image.

Like images, tables are defined in WML using the HTML specification. Tables consist of rows, which are made up of columns. The screen shown in Figure 9 was produced by the code below:

```
<p>
   <table columns = "2" >
     <tr>
       <td>Item:</td>
       <td>Cost:</td>
     </tr>
     <tr>
       <td> 1 </td>
       <td> $$27.80</td>
     </tr>
     <tr>
       <td> 2</td>
       <td> $$32.80</td>
     </tr>
     <tr>
       <td> 3 </td>
       <td> $$47.50</td>
     </tr>
   </table>
</p>
```

Notice the need to double the dollar sign symbol in order to display a dollar sign. Notice also that the table definition needs to be nested within a paragraph (`<p>..</p>`) element.

Although tables are useful for aligning information on the screen, the small size of the screen dictates a need for

Figure 9: A table of values.

Figure 10: Table with label.

restraint in displaying tables. Figure 10 displays the table obtained by replacing 1, 2, and 3 by labels. This small amount of information has consumed nearly all of the display space. This browser has chosen to wrap the text of column 1 entries. The browser will choose how to align and divide the table entries to fit the screen as well as possible. A developer should not invest a lot of time in fine tuning the display appearance. It will vary from one device to another.

User Input

User input is a challenge in the handheld wireless device world. There is no mouse and no full keyboard. Text input is awkward and slow. The easiest type of input for the user is selection from a list of options. The devices will all have some arrows for scrolling through lists and a numeric keypad for entering a number choice. There are occasions, however, when user input of text is necessary to provide a full-featured application. Whenever text input is required, the application should make the task as easy as possible for the user by allowing short inputs and using numbers wherever possible.

Input is accepted into a document by use of the `<input>` element. The syntax follows:

Introductory text

```
<input name = "variablename" size = "size"
  title = "label" format = "mask"
  maxlength = "numberchar"
  emptyok = "true | false" />
```

The introductory text describes what the application is expecting from this input. Size is the amount of space provided for this input field. Title is a label for this text. Format is the mask that will determine what is legal input from the user. If user input does not conform to the format specified, the input will be rejected. Maxlength is the maximum length of the text to be received from the user. The following code displays a box, accepts up to five characters from the user, makes the first character upper case and the others lower case, and will accept only letters or symbols.

```
<input name = "username" size = "5"
  title = "myname" format = "Aaaaa"
  maxlength = "7" emptyok = "false"/>
```

Legal entries for the mask include

`A` for any symbol or upper case letter, but no numbers;
`a` for any symbol or lower case letter, but no numbers;
`N` for a numeric character, but no letters or symbols;
`X` for a symbol, number, or upper case letter;
`x` for a symbol, number, or lower case letter;
`M` for a symbol, number, or upper case letter, changeable to lower case, defaults to upper case for the first letter; and
`m` for a symbol, number, or lower case letter, changeable to upper case, defaults to lower case for the first letter.

The format mask allows minor validity checking of user input. Although the wml code can do no processing, it can filter input so that only items matching the expected format are permitted. To check for reasonableness of values entered, the code must be supplemented with processing. Processing can be done by sending the data back to a server or by embedding script in the page itself. Sending data to a server involves the use of forms.

Forms

Forms in WML are closely related to forms in HTML. Users enter data that are transmitted to a server where the data become input to a process running on the server. The form contains text that describes the input needs of the application and explains to the user how to provide the needed information. Types of form elements for collecting data include

`Input elements`, with type either text or password, and a text input area. If the type is password, the data are not displayed during entry. If type is text, the characters are displayed as they are entered.

`Select elements` with their options.

`Input elements` without text input areas have either `type = checkbox` or `type = radio`.

`Input elements` corresponding to button controls have `type = submit` or `type = reset`.

`Input elements` with `type = hidden` are not displayed for the user to see, but provide data for processing at the server.

Clearly, the size of the display device will influence the design of effective forms. Generally, text input should be avoided unless essential. Descriptions and instructions must be kept short. Checkboxes, radio boxes, and selection lists will be easiest for the user and thus most likely to provide correct information to background processes. Most forms will require several screens to provide instruction to the user and adequate display of the information-gathering elements of the form. A significant challenge for the WAP application developer is to think in terms of small units of information presentation at any time, using the scarce real estate of the screen to best advantage.

Forms processing occurs in the same way that it does for traditional HTML forms. Form data are presented assembled and transmitted to a server process as input to a program. There is a difference in that a WAP proxy gateway will generally be in the communication path. This allows development of back end processing, whether through CGI scripts, Java server pages, or other technologies, without conscious thought about the restrictions associated with the small wireless devices. Output from the processing is then converted to a form suitable for display on the WAP client and forwarded through the gateway. The involvement of extra entities in the communication and service steps allows efficient development of applications to serve both standard and small-screen devices, but introduces an extra layer of security concern.

Security

There are four aspects to security in general network communications:

Privacy. Content is visible only to the intended recipient and both parties have confidence that privacy is protected. This is addressed with various levels of encryption. The degree of confidence required will be weighed against processing costs to determine the appropriate level of encryption to use.

Integrity. Content is not modified between leaving the sender and arriving at the recipient's device. Digital signatures allow document to be verified as being the same as was transmitted. A hash code is computed over the document and sent as part of the signature. If the hash code check on the recipient side does not produce the correct results, the document has been modified.

Authentication. The sender's identity can be verified with a very high degree of confidence. Passwords, authentication, and digital signatures identify the originator.

Nonrepudiation. The sender cannot later deny having sent the information. Digital signatures are the primary tools for binding the sender to the document or resource as sent.

The WAP architecture includes a wireless transport layer security specification, which includes a view of the wireless network access environment as shown in Figure 11. The figure shows both pull and push proxies. Network access is achieved in one of two modes: push or pull. Pull access is initiated by the client and causes information to be provided in response to a request. This is the familiar request/response scenario of most Web access. Push access is initiated by the sender. This involves a message delivered to the client without an explicit request from the client. Some examples of push technology are familiar: pagers, Short Message Service, and e-mail notification, when a user has signed up for the notification service. The user chooses this form of intrusion in order to remain aware of new activities or special offers. Additional push services are anticipated as the wireless Web develops.

End-to-end security is accomplished in the wired Web through secure socket layer (SSL) encoding. In that approach, a secure link is created end to end between the sender and the receiver. Intermediate processing units, such as routers, do not see the content of the message and only participate in routing the message from source to destination.

In the wireless Web, things are more complex. The client and server usually do not communicate directly, but rely on proxy or gateway machines to provide necessary translation and retransmission services. The proxy intervenes between the WAP-enabled wireless device and the TCP/IP and HTTP process-enabled server. Thus, security questions must include the degree of trust between the content provider and the gateway and between the client and the gateway, as well as between the client and the content provider.

WAP includes WTLS, wireless transport layer security. WTLS is used to provide secure service between a client device and its pull gateway. WTLS is used for server authentication. Client authentication, when required, is left to existing mechanisms. Nonrepudiation is left to the

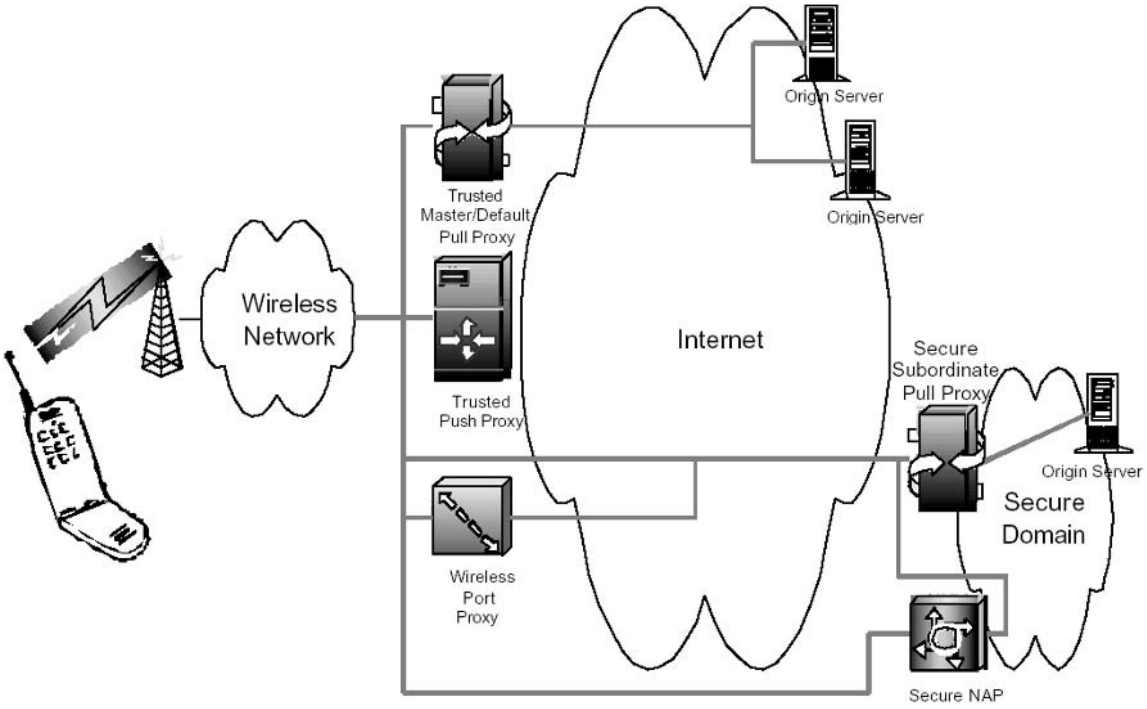


Figure 11: Wireless network access environment (Wireless Application Protocol, 2000). © 2000 Wireless Application Protocol Forum, Ltd. Reprinted with permission.

application and can rely on existing techniques. Control in the push environment is left to establishing trust between the client and the push proxy. The push proxy uses existing techniques to ensure security with data-providing servers.

The extra step involved is the use of a gateway between the client and the information provider. This introduces another factor into consideration of overall security. WTLS addresses security between the device and this pull gateway. Existing Web security mechanisms are used to provide security between the gateway and the service and/or data provider.

Design Considerations

Design considerations have figured in several parts of this chapter. The following points summarize the major issues.

Considering Speed

Bandwidth is still much lower for handheld wireless devices than for wired Web connections. As a result, bits that do not contribute to the user achieving a desired result will not be well regarded. Recent advances in design of engaging and entertaining Web pages will not serve this population. The application developer must focus on getting the message across in as few words, symbols, and images as possible.

There is another consideration related to speed, besides the limitation of bandwidth. It is important to consider where and how these devices are likely to be used. Often the device user is walking or taking time out from a meeting or other activity. He or she is not seated comfortably in his or her office or living room. The other constraints on these systems mean they are not generally the first choice for casual browsing. The user wants some information and wants it now. This is not the environment for enticing the visitor to explore the wonders of the site. A prompt and pointed response to a request will suit the user's needs and bring him or her back another time.

The Small Screen

The small screen has been mentioned a number of times and an example is shown in the illustrations. Although bandwidth and resolution on the screen will improve, the screen will not get larger, because smallness is a prized feature of many WAP-enabled devices. Some devices, such as PDAs, have larger screens than others. However, application developers must assume that the screen is small, often does not display color, and may be viewed in difficult lighting environments. Interaction must be designed for ease of use under these circumstances.

Navigation

Because of the small screen, the basic model of a "document" becomes a deck of cards. Each card contains a small amount of information and may contain a link to another card in the same deck or in a different deck. WAP provides all the power of the Web to reach sites and to initiate processes on remote services. However, the results of those server processes and the information on the Web pages will be constrained by the presentation environment. Each WAP client must have a way to interact with a specified set of events triggered by user actions. The way these are made available and the ease of user access will vary. An accept button and a button that goes back to the previous card are the most dependable constants. Presenting navigation options as links in a list allows the user to scroll through the choices and accept the most appropriate one.

LOOKING AHEAD

Expectations for Mobile Computing

It is hard not to think back to 1993 or 1994 when early Web pages were available and only a few people knew the basics of producing and deploying them. In the early years of the 2000s, the wireless Web is in a similar situation. Application development is necessary to drive demand, which in turn will drive more development. At the present time, the most successful implementation of a wireless Web environment is in Japan (Dornan, n.d.b.). Cell phone use is pervasive; the population is largely concentrated in cities where access to the cell infrastructure is nearly always available. Cell phones are used for Web access in ways that have not yet appeared even in the cell-phone-loving Scandinavian countries.

With success in Japan as an example, many analysts are predicting huge increases in the use of handheld wireless devices to access a growing list of services. Japan has a huge advantage not yet available in the rest of the world: the use of packet data transmission reduces the cost of having the phone on. Charges are based on the amount of data received, not on hours of connection. As a result, the phone is always on and access to service is always available. Only when that level of connectivity reaches the rest of the world will the promise of new services be realized.

The Role of the Wireless Application Protocol

In June 2002, the WAP Forum became part of a larger collaboration of industries and standards bodies called the Open Mobile Alliance. They describe the formation as follows:

> The foundation of the Open Mobile Alliance was created by consolidating the efforts of the supporters of the Open Mobile Architecture initiative and the WAP Forum. In addition, the Location Interoperability Forum (LIF), SyncML, MMS Interoperability Group (MMS-IOP), and Wireless Village, each focusing on mobile service enabler specifications, announced that they have signed a Memorandum of Understanding of their intent to consolidate with the Open Mobile Alliance. (Wireless Application Protocol Forum, 2001)

Thus, the work of the WAP Forum continues with a larger context and a much larger degree of commitment from the related industries. The role of these protocol efforts to join the industries that produce the client devices and those who produce the services and information sources will be crucial to the kind of open connectivity and roaming accessibility necessary to make the wireless Web as pervasive and important as the wired Web is today.

GLOSSARY

CDMA Code division multiple access. The basis of third-generation wireless transmission technologies.

Datagram A communication technique in which messages are broken into independent units that travel separately through network connections.

D-AMPS Digital advanced mobile phone systems.

Data push Access to data initiated by the provider, which pushes the data toward the consumer.

Gateway An intermediate device that links incompatible communication systems.

GSM Global system for mobile communications.

PDC Personal digital cellular.

Packet-based communication Another term for *datagram*.

Protocol stack A layered view of the collection of protocols required to accomplish a large task.

TDMA Time division multiple access.

WAE Wireless application environment.

WAP Wireless application protocol.

WCSS Wireless cascading style sheet.

WML Wireless markup language.

WBXML Compact binary representation of XML documents.

XHTML Extended hypertext meta language.

WSP Wireless session protocol.

WTP Wireless transmission protocol.

WTLS Wireless transport layer security.

CROSS REFERENCES

See *Extensible Markup Language (XML); HTML/XHTML (HyperText Markup Language/Extensible HyperText Markup Language); Mobile Devices and Protocols; Mobile Operating Systems and Applications; Wireless Communications Applications; Wireless Internet*.

REFERENCES

Digital mobile phone networks in Africa. Retrieved May 7, 2003, from http://www.cellular.co.za/gsm-africa.htm

Dornan, A. (n.d.a.). GSM and TDMA cellular networks. Retrieved May 7, 2003, from *NetworkMagazine.com*: http://www.networkmagazine.com/article/NMG20000517S0169

Dornan, A. (n.d.b.). WAP reaches the second generation. Retrieved May 7, 2003, from *NetworkMagazine.com*: http://www.networkmagazine.com/article/NMG20010823S0013

Evans, H., & Ashworth, P. (2001). *Getting started with WAP and WML*. Alameda, CA: Sybex.

ISO/IEC (1994). *Information technology—open systems interconnection—basic reference model: The basic model*. ISO/IEC 7498–1:1994.

Wireless Application Environment Specification Version 2.0 (2002, February 7). Retrieved May 7, 2003, from Open Mobile Alliance Ltd. Website: http://www1.wapforum.org/tech/documents/WAP-236-WAESpec-20020207-a.pdf

Wireless Application Protocol Wireless Session Protocol Specification (1999, November 5). Retrieved May 7, 2003, from WMLClub Website http://br.wmlclub.com/docs/especwap1.2/SPEC-WSP-19991105.pdf

Wireless Application Protocol Forum Ltd. (2001, July 12). Wireless Application Protocol Architecture Specification. Retrieved May 7, 2003, from Open Mobile Alliance Ltd. Website: http://www1.wapforum.org/tech/documents/WAP-210-WAPArch-20010712-a.pdf (Access

controlled from http://www.wapforum.org/what/technical.htm)

Wireless Application Protocol Forum Ltd. (2000, February 19). Wireless Transaction Protocol Specification. Retrieved May 7, 2003, from Open Mobile Alliance Ltd. Website: http://www1.wapforum.org/tech/documents/WAP-201-WTP-20000219-a.pdf (Access controlled from http://www.wapforum.org/what/technical.htm)

Wireless Application Protocol Forum Ltd. (2002, January). Wireless Application Protocol WAP 2.0 Technical White Paper. Retrieved May 7, 2003, from Open Mobile Alliance Ltd. Website: http://www.wapforum.org/what/WAPWhite_Paper1.pdf (Access controlled from http://www.wapforum.org/what/technical.htm)

Wireless Markup Language Version 2.0 (2001, September 11). Retrieved May 7, 2003, from Open Mobile Alliance Ltd. Website: http://www1.wapforum.org/tech/documents/WAP-238-WML-20010911-a.pdf (Access controlled from http://www.wapforum.org/what/technical.htm)

WAP Transport Layer End to End Security (2001, June 28). Retrieved May 7, 2003, from Open Mobile Alliance Ltd. Website: http://www1.wapforum.org/tech/documents/WAP-187-TransportE2ESec-20010628-a.pdf

Wireless Transport Layer Security V (2001, April 6). Retrieved May 7, 2003, from Open Mobile Alliance Ltd. Website: http://www1.wapforum.org/tech/documents/WAP-261-WTLS-20010406-a.pdf (Access controlled from http://www.wapforum.org/what/technical.htm)

FURTHER READING

Badrinath, B., Fox, A., Kleinrock, L., Popek, G., Reiher, P., & Satyanarayanan, M. (2000, December). A conceptual framework for network and client adaptation. *Mobile Networks and Applications, 5*(4) 221–231.

Bisdikian, C., Boamah, I., Castro, P., Misra, A., Rubas, J., Villoutreix, N., et al. (2002, September). Context and location: Intelligent pervasive middleware for context-based and localized telematics services. In *Proceedings of the Second International Workshop on Mobile Commerce* (pp. 15–24). New York: ACM Press.

Cohen, D., Herscovici, M., Petruschka, Y., Maarek, Y. S., & Soffer, A. (2002, May). Mobility and wireless access: Personalized pocket directories for mobile devices. In *Proceedings of the Eleventh International Conference on World Wide Web*. Retrieved May 7, 2003, from http://www2002.org/CDROM/refereed/92/index.html

Elaarag, H. (2002). Improving TCP performance over mobile networks. *ACM Computing Surveys (CSUR), 34*(3), 357–374.

Flynn, M., Pendlebury, D., Jones, C., Eldridge, M., & Lamming, M. (2000). The satchel system architecture: Mobile access to documents and services. *Mobile Networks and Applications, 5*(4), 243–258.

Fraternali, P., & Paolini, P. (2000). Model-driven development of Web applications: The AutoWeb system. *ACM Transactions on Information Systems (TOIS), 18*(4), 323–382.

Geng, X., Huang, Y., & Whinston, A. B. (2002). Defending wireless infrastructure against the challenge of DDoS attacks. *Mobile Networks and Applications, 7*(3), 213–223.

Hadjiefthymiades, S., Matthaiou, V., & Merakos, L. (2002). Supporting the WWW in wireless communications through mobile agents. *Mobile Networks and Applications, 7*(4), 305–313.

Jing, J., Helal, A. S., & Elmagarmid, A. (1999). Client–server computing in mobile environments. *ACM Computing Surveys (CSUR), 31*(2), 117–157.

Jones, C. E., Sivalingam, K. M., Agrawal, P., & Chen, J. C. (2001). A survey of energy efficient network protocols for wireless networks. *Wireless Networks, 7*(4), 343–358.

Joshi, A. (2000). On proxy agents, mobility, and Web access. *Mobile Networks and Applications, 5*(4), 233–241.

Marcus, A., & Chen, E. (2002). Designing the PDA of the future. *Interactions, 9*(1), 34–44.

Munson, J. P., & Gupta, V. K. (2002). Context and location: Location-based notification as a general-purpose service. In *Proceedings of the Second International Workshop on Mobile Commerce* (pp. 40–44). New York: ACM Press.

Olsson, D., & Nilsson, A. (2002). MEP: A media event platform. *Mobile Networks and Applications, 7*(3), 235–244.

Palen, L., & Salzman, M. (2002). Beyond the handset: Designing for wireless communications usability. *ACM Transactions on Computer–Human Interaction (TOCHI), 9*(2), 125–151.

Perry, M., O'Hara, K., Sellen, A., Brown, B., & Harper, R. (2001). Dealing with mobility: Understanding access anytime, anywhere. *ACM Transactions on Computer–Human Interaction (TOCHI), 8*(4), 323–347.

Phan, T., Huang, L., & Dulan, C. (2002). Challenge: Integrating mobile wireless devices into the computational grid. In *Proceedings of the Eighth Annual International Conference on Mobile Computing and Networking* (pp. 271–278). New York: ACM Press.

Samaras, G., & Panayiotou, C. (2002). Data and content: Personalized portals for the wireless user based on mobile agents. In *Proceedings of the Second International Workshop on Mobile Commerce* (pp 70–74). New York: ACM Press.

Shih, G., & Shim, S. S. Y. (2002). A service management framework for M-commerce applications. *Mobile Networks and Applications, 7*(3), 199–212.

Singhal, S., Bridgman, T., Suryanarayana, L., Mauney, D., Alvinen, J., Bevis, D., Chan, J., & Hild, S. (2001). *The Wireless Application Protocol*. New York: ACM Press.

Steinberg, J., & Pasquale, J. (2002). Mobility and wireless access: A Web middleware architecture for dynamic customization of content for wireless clients. In *Proceedings of the Eleventh International Conference on the World Wide Web*. Retrieved June 10, 2003, from http://www2002.org/CDRom/refereed/483/index.html

Varshney, U., & Vetter, R. (2002). Mobile commerce: Framework, applications and networking support. *Mobile Networks and Applications, 7*(3), 185–198.

Yoshimura, T., Yonemoto, Y., Ohya, T., Etoh, M., & Wee, S. (2002). Mobility and wireless access: Mobile streaming media CDN enabled by dynamic SMIL. In *Proceedings of the Eleventh International Conference on the World Wide Web*. Retrieved May 7, 2003, from http://www2002.org/CDROM/refereed/515/index.html

Wireless Communications Applications

Mohsen Guizani, *Western Michigan University*

INTRODUCTION

Statistics show that the level of demand for wireless communications in the past decade has exceeded expectations even when the cost of owning a mobile handset was at its highest. Some of the rational reasons are deregulation of telecommunications monopolies worldwide and adoption of internationally recognized standards by abandoning proprietary standards. Special tariff packages for mobile calls that cause prices to fall due to increased competition and aggressive marketing are being practiced.

It is evident that the wireless communications market has been evolving at an exponential rate since the 1990s and the numbers justify its growth today. One explanation for such growth patterns can be recognized by the replacement of analog by digital standards. Another well-accepted reason is a long-term objective to substitute wired communication with wireless communications technologies. Third world countries that lack a reliable wired-based communications infrastructure are adopting wireless communications technologies as a cheaper alternative. Consequently, wireless communications has proved to be one of the most profitable parts of the telecommunications sector worldwide.

At present, the majority of mobile communications users also own a fixed-link telephone. But, growth in demand for data and Internet-based services has been driving the wired-communication market over the past few years. However, if consumers want to incorporate data, voice, video, and Internet-based services in wireless communications, an infrastructure based on wireless asynchronous transfer mode (WATM) technologies can be used to integrate the best of the two worlds.

OSI MODEL FOR THE INTERNET

The open system interconnect (OSI) model divides communication into layers. Each layer fulfills certain tasks and makes different combinations for different applications. The different layers for the Internet models are the network interface layer, Internet layer (equivalent to the network layer on the OSI model and primarily includes IP), transport layer, and the application layer.

Internet Protocols (IP)

IP transports packets to the desired destination host on the network. IP is a connectionless protocol and is not aware of any sessions. Every packet is routed independently, and different parts of the same transmission might take a different route. Along the way the packet might be lost, corrupted, duplicated, or delivered out of sequence. If the underlying network is not capable of transmitting packets as large as those that higher layers try to get IP

to send, IP will fragment the packets in order to fit the network.

Transmission Control Protocol (TCP)

As a transport layer protocol, TCP ensures that data for which IP finds a destination host will be propagated to the right application. File transfer protocol (FTP) and hypertext transfer protocol (HTTP) applications might be running on the same host, and the TCP port number indicates which is the target application. TCP also ensures that packets of data are delivered reliably and in order. Packets that have traveled different routes in order to get to the same destination need to be assembled in the right order. Finally, TCP provides flow control functionality, which makes sure that the sender and receiver agree on a suitable speed of data delivery. Covering the lost of packets and the maintenance of flow control are the two features that affect wireless performance the most.

Hypertext Transfer Protocol (HTTP)

HTTP is a protocol that handles the transfer of Web pages and relies on TCP on the transport layer for the reliable, in-order delivery of individual packets. HTTP is a stateless protocol, and it has a number of different requests that a client can use from a server.

Each request is sent as regular ASCII text, and the server responds with an object. The object returned can either be a simple Web page or a MIME response containing several objects, such as images. When fetching each object of a Web page separately, a new HTTP request must be performed for every object. Because of the high latency of those networks, fetching each part of a Web page with a separate request might cause significant delay. While this situation might not matter much when you are using a desktop computer that has a broadband connection, it is a significant advantage when you are developing for wireless networks.

File Transfer Protocol (FTP)

Like HTTP, FTP is an application layer protocol. FTP is a connection-oriented file transfer between two hosts using TCP as transport. The only issue with FTP when running over wireless is greed. A client that fetches a file via FTP will start one TCP session and keep it throughout the transmission.

Before we get into the discussion of the different types of wireless networks/systems, Table 1 shows the comparisons between wireless and wireline network systems.

CELLULAR PHONE GENERATION

The first radiotelephone service was introduced in the United States at the end of the 1940s, and was meant to connect mobile users in cars to the public fixed network. In the 1960s, a new system launched by Bell Systems, called "improved mobile telephone service" (IMTS), brought many improvements like direct dialing and higher bandwidth. The first analog cellular systems were based on IMTS and developed in the late 1960s and early 1970s. The systems were "cellular" because coverage areas were split into smaller areas, which are served by a low-power transmitter and receiver.

First Generation (1G)

First generation (1G) phones are analog. It sends information as a continuously varying waveform, and can only be used for voice and have highly variable call quality. The disadvantage is it is very insecure. Snoopers can listen in on call or charge calls to another person's account. Two key improvements were made during the 1970s—the invention of the microprocessor and the digitization of the control link between the mobile phone and the cell site.

Second Generation (2G)

Second generation (2G) phones cover all speech into digital code, resulting in a clearer signal that can be encrypted for security. These systems digitized not only the control link but also the voice signal. The new system provided better quality and higher capacity at lower cost to consumers. It includes some kind of messaging, as well as support for Centrex-style services such as voice mail and caller ID. Most popular is GSM (global system for mobile communication), which can send data only at less than 10 Kbps (kilobits per second).

Second ½ Generation (2.5G)

Most operators are upgrading their 2G networks to higher data speed using fast modems. This term applies to technology such as WAP, which uses a compressed version of the Web to fit into a mobile phone's slow data rate and small screen. 2.5G networks, such as GPRS (global packet radio service), are already available in some parts of Europe.

Third Generation (3G)

Third generation (3G) systems promise faster communications services, including voice, fax, and Internet, anytime and anywhere with seamless global roaming. The ITU's IMT-2000 global standard for 3G is expected to enable innovative applications and services (e.g., multimedia entertainment, infotainment and location-based services, among others). Instead of a phone, many terminals will be small computers or PDAs (personal digital assistants). It will cover not only the connection between a mobile terminal and its base stations, called a WAN (wide area network), but also the LAN (local area network). The vision behind it is mobile multimedia to all at the price of today's fixed telephony.

Fourth Generation

Fourth generation networks are already in the labs, targeted for deployment beginning in 2010. It will provide data rates of up to 100 Mbps, enough for telepresence. This is a type of virtual reality, defined as full stimulation of all senses required to provide the illusion of actually being somewhere else—an illusion that cannot be distinguished from the real thing (Anderson, 2001; CDMA, n.d.).

Table 1 Comparison of Wireless and Wired Network Systems

	Wireless network	Wireline network
Transmission media	Radio, microwave, infrared and millimeter waves, lightwave/laser transmission	Twisted pair, coaxial cable, fiber optics
Standard	Wireless LAN: IEEE802.11	IEEE802.3 (based on Ethernet specification)
	Wireless MAN: IEEE802.16	IEEE 802.5
Reference model	IEEE 802 reference model: Physical layer Data link layer (MAC sublayer, LLC sublayer) Upper layer protocols	OSI reference model: Physical layer Date link layer (MAC sublayer, LLC sublayer) Network layer Transport layer Session layer Presentation Application layer
Physical layer	Wireless LAN: Infrared, FHSS, DSSS, OFDM, HR_DSSS, OFDM Wireless MAN: QPSK, QAM-16, QAM-64	Thick coax, Thin coax, Twisted pair, Fiber optics
Data rate	1–54 Mbps	Classical Ethernet: 10 Mbps Fast Ethernet: 100 Mbps Gigabit Ethernet: 1–10 Gbps Fiber channel: 100 Mbps–3.2 Gbps
MAC sublayer protocol	CSMA/CA (combines CSMA and MACAW)	CSMA/CD Random Contention
MAC frame format	3 classes of frame: data, control, management Data frame format: frame control, duration, address1, address2, address3, seq., address4, data, checksum	Preamble (8 bytes), destination address (6 bytes), source address (6 bytes), type (2 bytes), data, pad (0–46 bytes), checksum (4 bytes)
Logical link control (LLC)	Run on top of Ethernet and other IEEE802 protocols. Hides the differences between the various kinds of 802 networks by providing a single format and interface to the network layer	
Internet protocol	Mobile IP—Enable computer to maintain Internet connectivity while moving from one Internet attachment point to another Mobile IP basic ability—Discovery, registration, tunneling	Ipv4, Ipv6
Discovery	Use ICMP by adding the appropriate extensions to the ICMP header	ICMP
Routing algorithm	For mobile host: use foreign agent and home agent In ad hoc network: AODV (ad hoc on-demand distance vector) routing algorithm	Shortest path routing, distance vector routing, link state routing, etc.
Congestion control algorithm	Poor performance. Since transmission link is unreliable, packets are lost all the time. The algorithm uses the wrong assumption (that timeout is caused by congestion) for congestion control.	Good performance Assume that timeout is caused by congestion, not by packet loss
Transport layer protocol	Poor performance using TCP Solution: Indirect TCP; makes modification to the network layer code in the base station	TCP
Datagram protocol	Performance is far from perfect using UDP. Since the transmission link is unreliable WDP (wireless datagram protocol)	UDP

Table 1 continue

	Wireless network	Wireline network
Web application		
Protocol stack	WAP—based on existing Internet standard, such as IP, XML, HTML, HTTP	HTTP/TCP/IP
		—Application layer
	Wireless application environment (WAE)	—Session layer (like HTTP)
	Wireless session protocol (WSP)	—Transport layer (TCP)
	Wireless transaction protocol (WTP)	—Like Netscape SSL
	Wireless transport layer security (WTLS)	—UDP
	Wireless datagram protocol (WDP)	—Physical layer
	Bearer layer (GSM, CDMA, D-AMPS, GPRS)	
Makeup language	WML, C-HTML, HDML, XML, WMLScript	HTML, XML, JavaScript
ATM network		
Cell size	24-byte payload; header can be compressed and expanded to standard ATM	48-byte payload, 5-byte header
Media access control (MAC)	Problem: find suitable channel sharing/media access control technique at data link layer	CDMA
Data link layer	Needs to provide an error control, needs fragmentation, and reassembly	DHLC

CELLULAR NETWORKS

Cellular networks use many base stations, allowing them to cover a much wider area. The base stations are usually connected to each other using fiber optic cables to wireless links, then to external networks such as phone systems and the Internet.

A cell is the coverage area of a single base station. As a mobile phone moves through a network, it accesses services via the base station of whichever cell it is in. Radio waves radiate out from a transceiver. The further the mobile user is from the base station, the weaker the signal gets. The cell boundary is the limit where the mobile terminal can no longer send and receive reliably.

The overlap between cells also must be taken into account, as in some parts a user may be able to communicate with two or three base stations. One of the most important features of a mobile network is the ability for a user to move from one cell to another, originally known as roaming.

The process of switching a user from one cell to another while a call is in progress is called a handoff. Handoffs are very complex procedures because the base stations must calculate exactly when a user is crossing the cell boundary. This can take several seconds, so if users move too fast, their calls may be dropped.

The speed limit for analog systems is usually no more than 100 kph. Some digital systems can function at speed above 300 kph (on a high-speed train). No system can complete a handoff at the cruising speed of an airliner, one reason for not using a mobile phone on a plane (Collins, 2001; Dorman, 2001; *Evolution of a mobile market*, 2001).

Types of Handoff

There are three types of handoff:

Soft Handoff

Soft handoff ensures that a link is set up to the base station in the new cell before the old one is torn down. This system is very reliable. It should result in dropped calls only if the user is moving extremely fast or actually passes outside the cellular network. One problem is that a connection with two different base stations is very difficult to achieve. In most types of networks, adjacent base stations need to use different frequencies, while a phone can be tuned to only one frequency at a time.

Hard Handoff

Hard handoff requires that a phone break its connection with one base station before connecting to another. It is less reliable than soft handoff because a phone is not always able to establish a new connection. The new cell could already be full, or there may not be another cell available. A base station sometimes decides to make a handoff based on how far away a phone is, without considering whether another phone can pick it up. Hard handoff causes a noticeable break in conversation, which can be annoying for a user moving rapidly between small cells. To fix this problem, networks with microcell overlays tries to detect which users are moving and connects them to the main larger cell.

No Handoff

No handoff is very simple and relies on the mobile terminal actually making a new call once it has moved out of range on one transmitter. It is rare in traditional cellular networks because many mobile phone systems can take up to 30 s to set up a new call.

Effects of Frequency

The frequencies used by cell phones all lie within the UHF microwave band (same as used by TV transmission) ranging from 400 to 2000 MHz. Higher frequencies are more easily blocked by droplets of cloud in the atmosphere, which result in a shorter range. Networks based on these higher frequencies require smaller cells and more base stations. Lower frequencies initially seemed preferable

because fewer base stations mean lower cost (Dorman, 2001; Geier, 2001; Wasi, n.d.).

Voice Coding

All digital systems need to encode the analog waveforms of speech into a bitstream. The program used for this is called a codec (coder/decoder), often embedded within a special chip called a digital signal processor. Many different codecs are used in cell phones. The aim is to produce the lowest possible bit rate while maintaining acceptable sound quality. Because computing power is increasing continuously, new phones and networks are capable of using more advanced compression technology.

Waveform Coding

The simplest way to digitize a sound signal is to sample the waveform at regular intervals. This is called pulse code modulation (PCM) and is used by the DC and PSTN codec. A shorter interval will result in more accurate sampling but a higher bit rate.

Vocoding

Instead of sending an actual signal, a vocoder calculates how speech was produced and sends only the relevant pitch and tone information plus a description of the sender's mouth movements and vocal tract. A decoder then synthesizes a voice. Vocoders are used in a very low data rate situation. They can reach data rates as low as 1 Kbps.

Hybrid Codecs

Most codecs use a mixture of waveform and vocoding, based on synthesized speech with some PCM information. The precise bit rate depends on the quality of the sound and on how much processing power is available. For a mobile phone, the limiting factor is usually the need to compress and decompress in real time, using a battery-operated device. In any codec, there is a tradeoff between bandwidth and speech quality. Cellular networks typically use a codec between 5 and 13 Kbps.

PCS STANDARDS

PCS is any digital system that provides high-quality voice and narrowband data. It is the second generation of wireless technology, between analog cell phone and broadband mobile multimedia.

Three main categories of PCS are in use today. The first is the digital cell phones, which offer high-quality voice and limited data, the second category of PCS concentrates on data only, and the third category uses noncellular technologies in emergency services and large businesses. It works without a base station and supports automatic conference calls.

The largest category of PCS is used for digital cell phone networks. Although these were all designed primarily for voice, they can also support data transmission at varying rate. Most PCS systems are based on TDMA (time division multiple access), because FDMA (frequency division multiple access) is too wasteful of bandwidth, and CDMA (code division multiple access) had not yet been invented at the time they were standardized. The PCS system centers on spectral efficiency by how much capacity a system can squeeze out of its allocated frequencies.

GSM (Global System for Mobile Communication)

Global system for mobile communication is an open, nonproprietary system that is constantly evolving. One of its great strengths is the international roaming capability. It is used by more than half of all mobile phones. It gives consumers seamless and the same standardized contact ability in more than 170 countries. GSM satellite roaming has extended service access to areas where terrestrial coverage is not available.

What is technically distinctive about the technology is that GSM differs from first generation wireless systems in that it uses digital technology and time division multiple access transmission methods. Voice is digitally encoded via a unique encoder, which emulates the characteristics of human speech. This method of transmission permits a very efficient data rate/information content ratio.

The system operates on four different frequency bands. It was originally designed for frequencies around 900 MHz, to reuse the spectrum intended for Europe's analog TACS networks. It was later adapted to 1,800 MHz, licensed in Europe specifically for GSM, and then to 1,900 MHz, used in America for several different digital networks. Like other digital cellular technologies, GSM encodes data into waves using a form of phase modulation, a system that uses the different parts of a waveform to represent information. The precise type is known as GMSK, which achieves a symbol rate and data rate of 270.8 Kbps in each of its 200-KHz channels.

HSCSD (High-Speed Circuit-Switched Data)

HSCSD (high-speed circuit-switched data) is a very simple upgrade to GSM that gives each user more than one time slot in the multiplex. All HSCSD-capable networks use the enhanced data codec, so that each channel allows a rate of 14.4 Kbps. The standard allows up to four of these to be tied together, for a maximum of 57.6 Kbps. Because of the overheating problem, most HSCSD devices are asymmetric, allowing greater download than upload speed.

GPRS (General Packet Radio Service)

GPRS (general packet radio service) is the one most popular among operators of all the wireless Internet schemes. Designed for data, it promises to give every user a permanent and high-capacity connection to the Internet. GPRS represents a major step forward in mobile networks. Its key improvement is packet switching, which for most data application is more efficient than circuit switching. Packet switching uses bandwidth only when needed, freeing up gaps in the data stream for other uses. Because of this greater efficiency, the full specification calls for terminal scapable of using all eight time slots at once. The plan is that GPRS networks will eventually carry voice over packet, using a variable rate codec so the data can be transmitted during gaps in conversation.

D-AMPS (Digital Advanced Mobile Phone System)

D-AMPS (digital advanced mobile phone system) is designed to be compatible with older analog AMPS technology. It uses the 30-MHz frequency channel as AMPS, but divides each one up into three TDMA slots. D-AMPS uses the same-paired spectrum and structure as AMPS. The difference is that instead of sending a single FM radio transmission over a 30-MHz channel, it allows each one to be used by three simultaneous conversions. Operators of AMPS networks can selectively allocate channels to be either digital or analog, allowing the two systems to coexist. The modulation scheme is called differential quadrature phase shift keying (DQPSK).

PDC/JDC (Personal Digital Cellular/Japanese Digital Cellular)

Japan has deployed a system based on D-AMPS, but designed for backward compatibility with its own J-TACS analog system. Despite being used in only one country, it is the world's second most popular mobile standard (behind that of GSM).

J-TACS used channels of only 25 MHz, which means that some changes were needed to the D-AMPS system designed for 30 MHz. The same size time slots and modulation are used, resulting in a lower overall bit rate. However, the D-AMPS voice codecs and data rate of 9.6 Kbps can still be achieved, by missing out some of the error correction and protocol overhead. This makes the system less reliable, but this problem can be overcome by keeping the cells small so that every user has a relatively clear link to the base stations.

D-AMPS+

D-AMPS+ is a scaled down version of GPRS that is implemented in the United States. When ETSI designed GPRS, they hoped that it could be applied to every TDMA-based system, including D-AMPS and PDC. The principle is simple—use more slots in the multiplex to increase the capacity until a single user has the entire channel all the time. GSM has eight time slots, while D-AMPS has only three, and each of these already offers a lower data rate of 9.6 Kbps compared to GSM's 14.4 Kbps.

One problem is that to achieve a faster rate, terminals need to be able to transmit and receive at the same time. A standard D-AMPS phone transmits for one third of the time, receives for another third, and is idle for the remaining third. Without simultaneous transmission and receiving, there is only one spare time slot (compared to six in GSM), limiting speed to only 19.2 Kbps in one direction and 9.6 Kbps in the other.

CdmaOne

CdmaOne is the only 20th century system to use CDMA. It was developed by Qualcomm, and it has been standardized by the Telecommunications Industry Association as IS-95a. CDMA systems seem superficially simpler than those based on TDMA. They involve no slot of frame structure. Every phone just transmits and receives all the time, sending many duplicates of the same information to ensure that at least one gets through. The very high transmission rate is achieved by using two different phase modulation techniques, quadrature phase shift keying (QPSK) on the downlink and offset quadrature phase shift keying (OQPSK) on the uplink. The second type required more forward error correction, because individual phones cannot coordinate their transmission in the same way that base stations can.

One large disadvantage of CDMA systems is their power consumption. By transmitting everything 64 times, a CdmaOne phone would seem to drain its battery 64 times faster than necessary and cover its user with 64 times as much microware radiation. CdmaOne gets around this problem by carefully controlling the transmission power. The aim is to ensure that the signal strength at the base station is the same for every user, ensuring that all can be heard equally.

CdmaTwo

Phones based on CdmaOne already transmit and receive simultaneously and at very high data rates. Most of these data are redundant, but it doesn't have to be. A phone could use more than one of the available Walsh codes, multiplying its 16 Kbps capacity by any factor up to the number of calls per channel.

THIRD-GENERATION STANDARDS

Third generation systems are critical to the wireless Internet services often touted as the future of mobile communications. They will offer permanent access to the Web, interactive video, and voice quality that sounds more like a CD player than cellular phone. The term 3G was originally defined as any standard that provide mobile users with the performance of ISDN or better, at least 144 Kbps. Technologically, the increased capacity is found in part by using extra spectrum and in part by new modulation techniques that squeeze higher data rates from a given waveband. Vendors, operator, and regulators all accept that the move toward higher data rates and better services will be evolutionary. Standards must be backward compatible with their predecessors so that phones can maintain a connection while moving between cells based on the old and the new.

IMT-2000

The third generation system was first planned in 1992, when the ITU realized that mobile communications was playing an increasingly important role. An international study group predicted that mobile phones would rival fixed lines within 10 years. It began work on a project called FPLMTS (Future Public Land Mobile Telecommunications System), aiming to unite the world under a single standard. The acronym was awkward, so ITU adopted the name IMT-2000. The number 2000 represents the year 2000, which the ITU wanted to make globally available for the new technology. The number 2000 also means the data rate of 2000 Kbps and the frequencies in the 2000-MHz region.

The ITU's original definition of IMT-2000 concerned only the data rate. Three different rates were suggested,

each corresponding to a different type of ISDN, the standard for a carrier's core voice network. The first rate is 144 Kbps, which was the absolute minimum acceptable capacity. It is the same speed as a B-rate ISDN line, the type that can be deployed over ordinary telephone wires. B-rate ISDN makes up a large proportion of regular phone lines in some European countries. The second rate is 384 Kbps, which was the ideal capacity. It corresponds an H-rate ISDN channel, often used for videoconferencing. The third rate is 2 Mbps, which was the capacity that should be achievable inside a building. It corresponds to a European P-rate ISDN line, which is usually a fiber optic cable carrying up to 30 separate phone lines into an office switchboard.

Service Requirements

Just as none of 3G's 1992 founders foresaw mobile Web access, many of its ultimate applications may still not have been discovered. However, the industry is clear about its direction, and its goal is toward convergence: third generation's aim is to combine the Internet, telephones, and broadcast media into a single device. To achieve this, IMT-2000 systems have been designed with six broad classes of service in mind. Three of the service classes are already present to some extent on 2G networks, while three more are new and involve mobile multimedia. Here are the service classes in order of increasing data rate.

Voice—Even in the age of high-speed data this is still regarded as the "killer app" for the mobile market. Third generation will offer call quality at least as good as the fixed telephone network, possibly with higher quality available at extra cost.

Messaging—This is an extension to paging, combined with Internet e-mail. Unlike text-only messaging services built into some 2G systems, 3G will allow e-mail attachment.

Switched data—This includes faxing and dial-up access to corporate networks or the Internet. With always-on connection available, dial-up access ought to be obsolete.

Medium multimedia—This is likely to be the most popular 3G service. Its downstream data rate is ideal for Web surfing.

High multimedia—This can be used for very high-speed Internet access, as well as for high-definition video and CD-quality audio on demand.

Interactive high multimedia—This can be used for fairly high-quality videoconferencing or videophone, and for telepresence, a combination of videoconference and collaborative working.

Spectrums Requirements

In 1992, the ITU recommended that the entire world allocate the same frequencies to 3G services. This would enable easy global roaming, particularly if everyone was using the same IMT-2000 standard. Regardless of the location, the user could be sure that his or her mobile phone or data device would work. The only part of the 3G spectrums available worldwide is dedicated to satellite services. This problem is that while cellular 3G is only a year or two behind its original schedule, many analysts doubt that satellites capable of mobile 133-Kbps operation will ever get off the ground. Broadband satellites tend to need band now allocated to be released for cellular IMT-2000.

Compatibility

The ITU originally wanted a single global standard, but this has not been achieved. Instead, there are two main types of CDMA, and a third based on TDMA. The main reason for the dispute is compatibility with existing systems, which can be defined in three ways.

Directed upgrades—Network operators without a license for new spectrum need to deploy a system that is essentially just an improvement of what they already have, so that new phones will work with the older base stations and vice versa. Upgrades typically add packet switching and better modulation, but keep existing cell size and channel structure.

Roaming—In principle, a mobile terminal can be made to support any number of different systems, enabling them to be used worldwide.

Handover—Roaming is inconvenient for most users, as phones must be reset to use a different network. To make it easier, a 3G system can be built so that it actually hands over users to a 2G network as they move outside its coverage area.

The fundamental problem for IMT-2000 is that no single standard could both upgrade CdmaOne and hand over to GSM. This means that two very similar CDMA-based IMT-2000 standards are set for deployment, and which one is deployed depends entirely on the local 2G systems.

W-CDMA

Wideband CDMA (W-CDMA) is the system favored by most operators. It has been designed to allow handovers to GSM; however, GSM network cannot be upgraded to W-CDMA, although some components, such as GPRS backbone, can be reused.

The wideband designation refers to the channel bandwidth of 5 MHz. This is 4 times that of CdmaOne, and 25 times that of GSM. A wider bandwidth was chosen to allow higher data rates, though only in uncrowded area with very clear reception.

TD-CDMA (Time Division CDMA)

TD-CDMA (time division CDMA) sounds like a contradiction and is often referred to as a hybrid between TDMA and CDMA. The multiplex technique is CDMA, but time division duplexing is used to share a channel between uplink and downlink, making the most efficient use of spectrum as spare capacity not used for the uplink can be used for the downlink. TD-CDMA is not without cost in additional overhead, so W-CDMA operators tend to use it only for one channel, pairing the others off in the same way as other cellular systems.

UMTS (Universal Mobile Telecommunications Systems)

UMTS (universal mobile telecommunications systems) has been the planned European W-CDMA standard since 1996. Its development was spearheaded by the UMTS forum, an industry and government group charged with developing a successor to GSM. The UMTS Forum succeeded in developing a draft W-CDMA proposal that was compatible with GSM, but it underestimated worldwide demand for mobile communications. Before a full UMTS standard could be tested and ratified, the proposal was picked up by the Japanese. The first W-CDMA networks are being deployed in Japan by NTT DoCoMo and J-Phone (Kaaranen, 2001).

CDMA 2000

Of the 2G systems, only CdmaOne is already based on CDMA. This gives it a head start in the race to 3G, as operators are able to upgrade their existing networks with new software of modulation rather than building a new radio system. These upgrades are collectively known as CDMA 2000; all are backward compatible with existing IS-95 systems. Until mid-2000, the upgrade path for CdmaOne seemed clear. The end result was supposed to be a system named CDMA 2000 3XMC. It combines three channels, resulting in a wider band. Unfortunately, this system was not compatible with the form of W-CDMA. In 2000, Motorola and Nokia together launched a system called 1Xtreme, which they claimed can reach speeds similar to that of 3XMC, but using only one channel, and hence a third of the spectrum.

EDGE (Enhanced Data Rates for GSM Evolution)

EDGE (enhanced data rates for GSM evolution) is a technology based on TDMA. It was a planned upgrade for the GSM networks. The plan was that GSM operators would deploy it in their existing networks, while building UMTS to take advantage of the newly licensed IMT-2000 spectrum. Because UMTS can hand over calls to GSM, the two would even be compatible.

EDGE inherits almost all its main features from GSM and GPRS, including the eight-user TDMA structure and even the slot length of 0.577 ms. The only difference is the modulation scheme of 8-PSK, which triples the capacity compared to GSM.

EDGE could be applied to a regular GSM system, giving every user a landline quality voice or data connection, but it is such a major upgrade that every operator who installs it will also use GPRS and HSCSD. Because 8-PSK is more susceptible to errors than GSM, EDGE has nine different Modulation and Coding Schemes, each designed for a different quality connection.

To make EDGE deployment easier for a D-AMPS operator, the UWCC (Universal Wireless Communications Consortium) has defined a simplified stand called EDGE Compact. This can be used only for data, not voice, and so omits many of the control channels found in the full-scale system, which in the context of EDGE compact is referred to as EDGE Classic.

Finally, Table 2 summarizes the evolution of wireless systems.

MOBILE DATA SERVICES

Messaging

Every digital, and even some analog, mobile phone systems already incorporate some form of messaging. This allows subscribers to receive and sometimes send short text messages. This is essentially the same as paging, but with the data appearing on a mobile phone instead of a separate page.

In theory, messaging service should allow people to receive e-mail through their mobile phone and to dispense with pagers altogether. However, in practice, neither of these objectives has been achieved. Mobile phone operators are not geared up to deal with the type of services offered by paging companies. Most of the messages sent to a pager originate on the regular telephone network. To page someone, the user calls an operator, who transcribes the message and transmits it. Operators have made more progress in interconnecting their own messaging networks so that customers of one operator can send messages to those of another.

A digital cellular phone uses three types of messaging services:

SMS (short message service)—Only messaging standard to have achieved widespread acceptance. Began as part of the original GSM specification, but has since spread to all the other digital systems. Limited versions must be standardized for the AMPS and NMT analog systems.

CBS (cell broadcast service)—If the same information needs to be sent to many different users, broadcasting is more efficient than transmitting a separate transmission. Each message is known as a page and can be only 93 bytes long. Up to 15 pages can be concatenated together. CBS has not been widely deployed because it offers operators no way to charge for the services.

USSD (unstructured supplementary services data)—This service uses the control channel and can operate while a phone in use. The message is longer than that of SMS, with a maximum of 182 bytes. One advantage of USSD is it is connection-oriented. The network establishes a connection with the phone before sending any data.

The Wireless Web

Transferring the Web to cell phones presents several changes, including variable latency and input device design. Two issues that have preoccupied the mobile industry are low capacity and small screens. Most existing cell phone systems allow data speed of only 13.3 Kbps. Even more advanced technologies will push this up to 56 Kbps. Because of the small screen, many cellphones can display only a few lines of monochrome text, whereas a PDA has larger displays, but it still has limits. It is not possible to squeeze a desktop-sized screen into the palm of a user's hand.

Table 2 Evolution of Wireless Systems

1G systems
Established in late 1970s
Based on analog technology
Aimed at providing voice telephony service
Major technologies: AMPS, TACS, ETACS, NMT, JTACS, NTACS, NTT, C-450, RMTS, Radiocom
2G systems
Established in early 1990s
Based on digital and PCS technology
Aimed at providing a better spectral efficiency, a more robust communication, voice privacy, and authentication capabilities
Major technologies: GSM, TIA/EIA/IS-136 or D-AMPS, JDC, PHS, TIA/EIA/IS-95A, IS-54, PACS, CT-2, DECT
2.5G systems
Based on 2G Systems
Aimed at providing the 2G systems with a better data rate capability
Major technologies: GPRS, IS-95B, HDR, GSM/EDGE
3G systems
First planned in 1992, embodied by IMT-2000
Based on digital and PCS technologies
Supports hybrid air interfaces
Aimed at supporting multimedia services
Major technologies: UTRA, W-CDMA, UWC-136, IMT-2000, CDMA 2000, UMTS
4G systems
Developed in the labs, targeted for deployment beginning in 2010
Aimed at providing data rates of up to 100 Mbps

C-HTML (Compact HTML)

C-HTML (compact HTML) is a simplified version of HTML. As the Web became more focused on powerful computers with high-bandwidth connections, it was decided to create a special version of HTML for devices with limited computing power. It has only the core textual display language and simple graphics. The advantage of C-HTML is that it displays perfectly normal on any regular Internet browser.

HDML (Handheld Device Markup Language)

HDML (handheld device markup language) is a more radical departure from HTML. HDML requires that a site avoid the use of tables, frames, flashing lights, and other complicated features. HDML replaces the familiar concept of Web pages with two new test layout metaphors called card and decks. A card is a single-user interaction, which would equate to a page in HTML. A single HDML file is called a deck because it can include many cards.

XML (Extensible Markup Language)

XML (extensible markup language) is a code that lets Web authors define their own tags. This means that it can be used for any type of device that becomes Internet-enabled. XML provides more flexibility. Unlike HTML, it is not limited to the description built in the language.

WIRELESS LANs

The emergence and continual growth of wireless LANs are being driven by the need to lower the costs associated with network infrastructures and to support mobile networking applications that offer gains in process efficiency, accuracy, and lower business costs (Wood, n.d.).

The implementation of wireless networks offers many tangible cost savings when performing installations in difficult-to-wire areas. If rivers, freeways, or other obstacles separate buildings you want to connect, a wireless solution may be much more economical than installing physical cables or leasing communication circuits. The deployment of wireless networking in these situations costs thousands of dollars, but it will result in a definite cost savings in the long run.

A problem inherent to wired networks is downtime due to cable faults. With wired networks, a user might accidentally break the network connector when trying to disconnect a PC from the network to move it to a different location. Wires and connectors can easily break though either misuse and/or normal use. An advantage of wireless networking, results from the use of less cables. This reduces the downtime of the network and the costs associated with replacing cables.

The installation of cabling is often a time-consuming activity. For LANs, installers must pull twisted-pair wires or optical fibers above the ceiling and drop cables through walls to network outlets that they must affix to the wall. These tasks can take days or weeks. The deployment of wireless networks greatly reduces the need for cable installations, making the network available for use much sooner.

Unfortunately, wireless LANs suffer from a fatal flaw—they are slow. The result is that wireless LANs have

remained a niche, used only in environments such as warehouses, where mobility is essential.

Wireless LAN Standards

The main problem faced by early wireless LANs was that there were no real standards. Each company produced its own proprietarily systems. Many simply chose not to buy a wireless LAN at all, for fear of being locked into a single vendor's technology. Standards began to emerge in the late 1990s, but potential customers were again spoiled for choice. Most wireless LAN standards are spawned by either the IEEE or ETSI. Between them, they have produced at least six incompatible standards.

Table 3 gives a summary of the wireless LAN standards available.

IEEE 802.11

IEEE 802.11 was the first wireless LAN standard to be defined. It uses the same switching protocols as wired Ethernet, but gave up wire in favor of ISM radio. Just to make things more confusing, it is actually two standards in one, each employing an incompatible type of spread spectrum.

The simplest version of 802.11 uses a frequency-hopping spread spectrum, rapidly cycling between frequencies many times each second. It has 70 different frequencies to choose from, so while some may be blocked, another is clear. If information doesn't get through, it is resent. A more complex version uses a direct sequence spread spectrum (DSSS). It transmits on all frequencies simultaneously. This increases the data rate, but also uses more powers.

IEEE 802.11b

IEEE 802.11b is based on a DSSS version of IEEE 802.11. It uses a better modulation technique to increase capacity up to a maximum of 11 Mbps. As well as a large increase, this speed also pushes to match the original Ethernet standard.

IEEE 802.11a

IEEE 802.11a reaches a speed of 54 Mbps by abandoning ISM and spread spectrum. It uses the U-NII band and a technique called coded orthogonal frequency division multiplexing (OFDM), which is designed to minimize interference caused by a signal reflecting off walls.

HiperLan 1

HiperLan 1 was the standard proposed by ETSI in 1992, when it first recognized the need for high-speed wireless LANs. Much of it is based on Ethernet, although its radio access technology was taken straight from GSM.

HiperLan 2

HiperLan 2 uses the same spectrum as HiperLan 1, but is otherwise much closer to IEEE 802.11a. It has a maximum data rate of 54 Mbps, which is achieved by using the same coded OFDM technology.

HomeRF

HomeRF is designed for home networking. It is based on the original FHSS version of 802.11, and designed primarily for price rather than speed. Its major innovation is direct support for telephony. The other standards are all designed for data only, needing additional software to be able to carry voice.

Access Points

Wireless LAN systems can all be used for ad hoc networking between two or more users who happen to have a card installed. Areas such as offices or homes can also be fitted with access points, which both extend the range of the system and enable it to link to ordinary LANs or the Internet.

Private users can create their own miniature cellular network by setting up several access points. The entire standard includes a handover mechanism, similar to those of public cellular networks. The IEEE 802.11 family inherits the soft handoff system from CDMA cellular, meaning that the mobile unit tries to form a link with a new access point before it disconnects from the previous one. Hyper LAN systems use a hard handover similar to GMS, which means that they must disconnect from one access point before reconnecting to another. This is less reliable and results in a short interruption in connectivity.

Access points can take two forms, hubs and switches. A hub is the simplest, simply rebroadcasting everything it receives. A switch is more discriminating, sending transmission only to smaller subgroups known as segments. They improve network performance, as the total capacity is shared per segment.

On a wired network, there is no limit to the number of segments. A switch could put every user on its own segment, giving each access to the full capacity all the time. Wireless networks cannot do this, because each segment needs its own spectrum. However, they could make it much easier to reallocate machines between different segments, a task that usually requires rewiring. A wireless network can do this at the push of a button, or even automatically.

WIRELESS ATM

ATM has been advocated as an important technology for the wide area interconnection of heterogeneous networks. In ATM networks, the data are divided into small, fixed length units called cells. The cell is 53 bytes. Each cell contains a 5-byte header that comprises identification, control priority, and routing information. The rest of the 48 bytes are the actual data. ATM does not provide any error detection operations on the user payload inside the cell, and also provides no retransmission services, and only few operations are performed on the small header (Guizani & Rayes, 1999).

There are some important factors that might help ATM make it successfully in the wireless world. These include flexible bandwidth allocation and service-type selection for a range of applications, efficient multiplexing of traffic from bursty data/multimedia sources, end-to-end

Table 3 Wireless LAN Standards

Standard		Operating frequency	Data rate	Modulation	Functionality
IEEE 802.11	Infrared	Operating at a wavelength between 850 and 950 mm	Omni-directional; range is up to 20 m; 1 Mbps, and 2-Mbps data rate	16-PPM (pulse position modulation)	Distribution service: association, disassociation, reassociation, distribution, integration
	FHSS	Operating in the 2.4-GHz ISM band	Make use of a multiple channels, with single hopping from one channel to another based on a pseudonoise sequence. Data rate is 1 Mbps and allows for optional 2Mbps	2-level gaussian frequency shift keying (GFSK) modulation for 1 Mbps; 4-level GFSK for 2 Mbps	Station service: authentication, deauthentication, privacy, data delivery
	DSSS	Operating in the 2.4-GHz ISM band	Use up to 7 channels, each with 1- and 2-Mbps data rates	Data rate of 1 Mbps for DBPSK and 2 Mbps for DQPSK	Designed to be mobile Ethernet
IEEE 802.11b		DSSS schema	5–11 Mbps	CCK	
IEEE 802.11a		Operating in 5-GHz band, support OFDM.	6, 9, 12, 18, 34, 36, 48, 54 Mbps	BPSK, QPSK, 16-QAM, 64-QAM	
IEEE 802.16		Operating at 10–66 GHz frequency range		QAM-64, QAM-16, QPSK	Designed to be wireless, but stationary. Supports heavy-duty multimedia usage.
IEEE 802.16a		Supports OFDM in the frequency range 2–11 GHz			
IEEE 802.16b		Operates in 5-GHz ISM band			
HiperLan 1		Provides up to 20-Mbps data rate in the 5-GHz range of the radiofrequency (RF) spectrum.	Can only use one channel,	High-transmission rate: 24 Mbps, GMSK. Low-transmission rate: 1.47 Mbps, FSK	For ad hoc networking of portable device. Does not control or guarantee QoS on wireless link
HiperLan 2		Supports OFDM	Up to 54 Mbps data rate for short-range (up to 150 m) communication	BPSK, QPSK, QAM-16, QAM-64	Support asynchronous data and time critical services, efficient QoS, ATM cells. Typical application include offices, homes, exhibition halls, airports, train stations, and so on

provisioning of broadband services over wireless and wireline networks, and ease of interfacing with the wired networks systems that will form the telecommunications backbone (Wasi, n.d.).

Adoption of an ATM compatible fixed-length cell-relay format for PCN will result in a relatively transparent interface to an ATM backbone. By using ATM switching for intercell traffic, the crucial problem of developing a new backbone network with sufficient throughput to support intercommunication among large numbers of small cells is avoided. It is noted that for PCN micro- and picocells with relatively low traffic volumes, rather than direct

connection to an ATM switch, it may be appropriate to use a lower cost shared media approach (such as TDM passive optical network or IEEE 802.6 optical bus) to interconnect several base stations.

Wireless ATM Key Issues

Architecture

The wireless ATM architecture is composed of a large number of small transmission cells, called picocells. Each picocell is served by a base station. All the base stations in the network are connected via the wired ATM network. The use of ATM switching for intercell traffic also avoids the crucial problem of developing a new backbone network with sufficient throughput to support intercommunication among a large number of small cells. To avoid hard boundaries between picocells, the base stations can operate on the same frequency.

Reducing the size of the picocells has major advantages in mitigating some of the major problems associated with in-building wireless LANs. The main difficulties encountered are the delay due to multipath effects and the lack of a line-of-sight path resulting in high attenuation. Picocells can also have some drawbacks as compared to larger cells. There are a small number of mobiles within the range of any base station, so a base station's cost and connectivity are critical. Once the cell size is reduced, the handover rate also increases. By using the same frequency, no handover will be required at the physical layer. The small cell sizes also give the flexibility of reusing the same frequency, thus avoiding the problem of running out of bandwidth.

The mobile units in the cell communicate with only the base station serving that particular cell, and not with other mobile units. The basic role of the base station is interconnection between the LAN or WAN and the wireless subnets. It is also used to transfer packets and convert them to the wired ATM network from the mobile units. In traditional mobile networks, transmission cells are "colored" using frequency division multiplexing or code division multiplexing to prevent interference between cells. Coloring is wasteful of bandwidth since in order to be successful there must be areas between reuse of the color in which it is idle. These inactive areas could potentially be used for transmission (Wasi, n.d.).

Cell Size

The ATM cell size of 53 bytes is designed for 64 Kbps or higher, which may be too large for some wireless LANs due to low speed and high error rates. Therefore, wireless LANs may use a 16- or 24-byte payload. The ATM header can also be compressed and be expanded to standard ATM at the base station. An example of ATM header compression is to use 2 bytes containing 12-bit VCI (virtual channel identifier) and a 4-bit control (payload type, cell loss priority, etc.).

One of the cell formats is to have a compatible payload size and addressing scheme, which should be different from the standard ATM cell format. Mobility should be as transparent as possible to the end-points, and therefore the VCIs used by the end-points should not change during handover. The allocation of the VCI should remain valid as the mobile moves through different picocells within the same domain. The translation of the VCIs should be as simple as possible due to movement between domains. This can be done by splitting the VCI space into a number of fields like domain identifier, mobile identifier, base station identifier, and virtual circuit number. A 16-bit CRC is also used to detect bit errors, due to the high error rate of mobile networks.

Physical Layer

The basic design issue for next generation private communication network (PCN) is the selection of modulation methods and a set of bit rates. A bit rate in the range of 5–10 Mpbs can be achieved using the existing wireless technologies, in a picocellular environment. Thus, with the exception of HDTV, most other ATM applications can be supported. The preferred technique may actually vary with the specific PCN application scenario to be addressed, so that it is likely that both TDMA and CDMA solutions will coexist. CDMA provides an efficient integrated solution for frequency reuse and multiple access, and can typically achieve a net bandwidth efficiency 2–4 times that of comparable narrowband approaches. However, a major weakness of CDMA for multiservice PCN is that for a given system bandwidth, spectrum spreading limits the peak user data rate to a relatively low value.

Narrowband (TDMA) can be used to achieve high bit rates. In a picocellular environment, using the narrow band approach can achieve a bit rate in the range of 8–16 Mbps. Overall, it should be possible for macro- (5–10 km), micro- (0.5 km), and pico- (100 m) cells to support baud rates on the order of 0.1–0.25, 0.5–1.5, and 2–4 Msym/s, respectively. These rates should be sufficient enough to accommodate many of the broadband services (Wasi, n.d.).

Media Access Control (MAC)

One of the major problems of wireless ATM is finding a suitable channel sharing/media access control technique at the data link layer. Shared media access leads to poor quantitative performance in wireless networks. When spread spectrum modulation is used, CDMA is the de facto mode of operation. Performance results from earlier studies shows that packet CDMA can achieve good traffic multiplexing efficiency and performance for CBR (constant bit rate), VBR (variable bit rate), and low-speed interactive data services, but CDMA is limited to less than or equal to 1 Mbps at higher speeds.

Data Link Layer

Wireless ATM needs a custom transparent data link layer protocol. A custom data link protocol is needed due to a high error rate and different packet sizes of wireless ATM. Wireless ATM may use a 16- or 24-byte payload, as 53 bytes may be too long for some wireless ATMs. The data link protocol may contain a service-type definition, error control, segmentation and reassembly, and handoff support.

A service-type field is needed so as to indicate whether a packet is of type supervisory/control, CBR, VBR ABR, etc. The service-type field simplifies base station protocol processing. Wireless ATMs should provide an error

control due to high noise interference and poor physical level characteristics of the wireless medium. This is achieved using a PCN packet sequence number filed in the header along with a standard 2-byte CRC frame check sequence trailer. HDLC-style retransmission can be used for connectionless data.

Since wireless ATM may use 16- or 24-byte cells, segmentation and reassembly is required. As a handoff is an important characteristic of wireless networks, a soft handoff without any data loss is vital and should be transparent. This can be implemented by using bits in the header, which indicates PDUs before and after the handoff.

FUTURE PHONES

The wireless Internet industry is unsure about what type of devices people will want, or more accurately, what they can persuade consumers to buy. The consensus is around smart phones. There are two competing philosophies about the future of mobile devices:

Personal area network—This assumes a modular system, whereby people continue to carry a plethora of devices as they do now. The difference is that all are linked together by a short-range radio system such as Bluetooth, with the cellphone acting as central hub and router.

Integrated device—This assumes that people will want to carry only one device but vary it according to the situation. A user might want to travel light and use only voice communication. At other times, he/she will want to use a full feature computer with an internal high-speed wireless data connection. The mobile network automatically detects which terminal is in use and routes calls accordingly.

Wearable Computers

Anything smaller than a telephone will need a headset containing an earpiece and a microphone. For video on the move, several vendors have proposed putting tiny screens inside sunglasses or contact lenses. The latter has the potential to immerse users completely in virtual reality. The problem with any small mobile terminal is that its battery will quickly run down. Shoes can be used to contain a reasonable-sized battery.

Smart Phones

A smart phone is a mobile phone with some extra computer-type functions, like a WAP browser. Besides mobile data capabilities, smart phone features include location-based services and sophisticated address books that can interface with a computer.

Tablets

A tablet PDA is one that doesn't have a keyboard and relies on a touch-sensitive screen for input.

Clamshell

Before the arrival of large, touch-sensitive screens, nearly all PDAs use the same clamshell design as laptops.

Subnotebooks

Subnotebooks sit on the borderline between PDAs and full-fledged laptop computers. The first subnotebook was Atari Portfolios. It was designed to be IBM compatible, running MS-DOS.

CONCLUSION

The Web is going wireless. It is expected that by 2003, more people will access the Internet via mobile phones than through computers. Most existing mobile phone systems are 2G. Third generation terminals will support high-speed data services. Most 2G systems, including CSM, are based on TDMA. CDMA is a more advanced technology, but does not have the installed base of GSM. It is used as the basis for 3G systems. TDMA usually allows each user to transmit or receive only part of the time. There are three main 3G systems: W-CDMA, CDMA 2000, and EDGE. They are collectively known as IMT-2000 and will offer packet-switched data at rates exceeding 384 Kbps. W-CDMA, known as UMTS, is designed to be backward compatible with GSM, and requires new spectrum. CDMA 2000 is an upgrade to CdmaOne; EDGE is an upgrade to GSM and is compatible with other TDMA systems. The wireless Web opens up many new business opportunities, the most important of which use location-based technology. Phones and computers are converging, but no one knows exactly what the result will be. Computers may eventually become wearable, but several problems must be overcome. These include battery life and user input.

GLOSSARY

ATM Asynchronous transfer mode.
CBS Cell broadcast service.
CDMA Code division multiple access.
C-HTML Compact HTML.
CSMA/CA Carrier sense multiple access/collision detection.
D-AMPS Digital advanced mobile phone system.
EDGE Enhance date rates for GSM evolution.
FDMA Frequency division multiple access.
FTP File transfer protocol.
GPRS General packet radio service.
GPRS Global packet radio service.
GSM Global system for mobile communication.
HDML Handheld device markup language.
HiperLan 1 Wireless LAN standard based mainly upon the Ethernet, although its radio access technology was taken straight from GSM, and createed due to the need for high-speed connection.
HiperLan 2 Uses the same spectrum as HiperLan 1, but is otherwise much closer to IEEE 802.11a.
HomeRF Designed for home networking and based on the original FHSS version of 802.11.
HSCSD High-speed circuit-switched data.

HTTP Hypertext transfer protocol.
IEEE 802.11 The first wireless LAN standard to be defined.
IEEE 802.11a Reaches a speed of 54 Mbps by abandoning ISM and spread spectrum.
IEEE 802.11b Based on the DSSS version of IEEE 802.11 and uses a better modulation technique to increase capacity up to a maximum of 11 Mbps.
IMTS Improved mobile telephone service.
IP Internet protocol.
IPV4, IPV6 Internet protocol version 4, Internet protocol version 6.
LAN Local area network.
LLC Logical link control.
OSI Open system interconnect.
PCM Pulse code modulation.
PDC/JDC Personal digital cellular/Japanese digital cellular.
SMS Short-message service.
TCP Transmission control protocol.
TDMA Time division multiple access.
UMTS Universal mobile telecommunications systems.
USSD Unstructured supplementary services data.
UWCC Universal Wireless Communications Consortium.
WAE Wireless application environment.
WAN Wide area network.
W-CDMA Wideband CDMA.
WAP Wireless access protocol.
WDP Wireless datagram protocol.
WSP Wireless session protocol.
WTLS Wireless transport layer security.
WTP Wireless transaction protocol.
XML Extensible markup language.

CROSS REFERENCES

See *Extensible Markup Language (XML); Mobile Devices and Protocols; Mobile Operating Systems and Applications; TCP/IP Suite; Wireless Application Protocol (WAP); Wireless Internet*.

REFERENCES

Anderson, C. (2001). *GPRS and 3G wireless applications*. New York: Wiley.

CDMA—Next Generation Wireless: 3G. (n.d.). Retrieved September 2002 from http://www.engineeringlab.com/3gwcdmagsm.html

Collins, D. (2001). *3G wireless networks*. New York: McGraw-Hill.

Dornan, A. (2001). *The essential guide to wireless communications application*. Englewood Cliffs, NJ: Prentice Hall.

Evolution of the mobile market. (2002, June 4). Retrieved September 2002 from http://www.itu.int/osg/spu/ni/3G/technology/

Geier, J. (2001). *Wireless LANS: Implementing high performance IEEE 802.11 networks* (2nd ed.). Indianapolis, IN: Sams.

Guizani, M., & Rayes, A. (1999). *Designing ATM switching and networks*. New York: McGraw-Hill.

Kaaranen, H. (2001). *UMTS networks: Architecture, mobility, and services*. New York: Wiley.

Wasi, A. S. (n.d.). *Wireless ATM*. Retrieved September 2002 from http://www.cis.ohio-state.edu/~jain/cis788–95/wireless_atm/index.html

Wood, J. B. (n.d.). *The wireless LANs page*. Retrieved Septemper 2002 from http://www.cis.ohio-state.edu/~jain/cis788–95/ wireless_lan/

Wireless Internet

Magda El Zarki, *University of California—Irvine*
Geert Heijenk, *University of Twente, The Netherlands*
Kenneth S. Lee, *University of Pennsylvania*

INTRODUCTION

During the past decade, we have experienced a tremendous growth and popularity of two communication technologies: cellular telephony and the Internet. Digital systems of the second-generation cellular networks have driven the cost down to an affordable level, and the public has been very receptive to the newfangled and revolutionary possibility of being able to communicate with others with reasonable service quality regardless of their location. Thus far, the primary mode of communication over such an infrastructure has been voice. Limited data applications over cellular and other wireless networks have simultaneously made some inroads, with paging and messaging services, including electronic mail (e-mail), having captured some customer interest, but still with very rudimentary forms of Internet services provided.

Subscriber Growth for Both Internet and Cellular Services—Expectations for the Mobile Internet

Cellular networks were introduced in the 1980s, and were initially designed to offer voice service. The first generation of cellular phones were built on an analog platform, but the current second generation is digital. In the United States, about 50% of households, over 130 million subscribers, now use cellular phones. Future wireless networks are expected to transition to an architecture that more closely resembles the Internet. Based on a usage/marketing study conducted by Qualcomm (see Figure 1), wireless services are expected to grow, justifying the need for higher bit rate services and a move to an architecture that is more data friendly.

The growth of the Internet has been equally strong, driven by the ability to access vast amounts and varieties of data and other resources whenever, and increasingly wherever, it is convenient for the end-user. Unlike the voice-centric cellular network, the strength of the Internet lies in its flexibility, in terms of the type and the number of different applications that can be used, and not in the optimization of a single application. Up to now, access to the Internet has mostly been through fixed terminals or computers rather than wireless devices.

The Internet, created in 1969 by the Department of Defense Advanced Research Projects Agency, was designed to offer file transfer service. Taking advantage of openly published rules of operation and freely distributed software, many research and educational institutions attached their computers to the Internet during the 1970s. The network has largely blossomed, however, due to the introduction of personal computers during the 1980s and the development of the World Wide Web in the 1990s. The number of people worldwide using the Internet has approximately doubled every year since the early 1980s, and is currently estimated to be over 660 million. The Internet is now in the process of transitioning toward an architecture that can better support real-time applications such as voice and video.

The Case for Mobility and Wireless Links in the Internet

Both the first- and second-generation wireless networks were designed and built for a world where voice was the driving application. However, the proliferation of the Internet has changed the expectation of the mobile user.

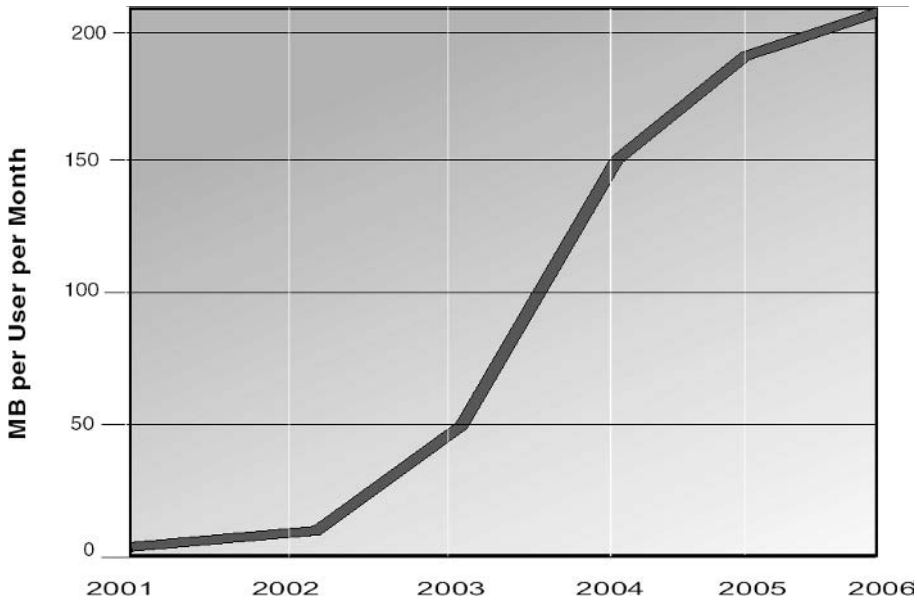


Figure 1: Wireless mobile data forecast. Source: Qualcomm (forecast date: 2001).

Mobile users are beginning to demand from their mobile wireless devices the types of services to which they have become accustomed on the Internet, including Web access, e-mail, and audio and video streaming. The forecast of the increasing number of users of wireless services and of the Internet is illustrated in Tables 1 and 2. It is natural to conclude that wireless devices will become an important access device to the Internet. However, Internet applications have vastly different characteristics and requirements than voice telephony, and thus are not efficiently supported by the current cellular wireless systems, which had been designed primarily for voice telephony application.

Use of Internet Protocol (IP) Technology in the Cellular Network

Since current cellular wireless networks have been designed and optimized for circuit-switched voice communications, the transformation of the radio access network (RAN) has been difficult. The RAN identifies the portion of the wireless network that handles the radio frequencies (RF) connections, both the transfer of the voice or information and the signaling needed to manage the so-called air interface, including the data synchronization aspects of the transmission. Only recently has there been work to utilize Internet protocol (IP) in a network that connects the base stations (BSs). To promote application independence and to decrease costs for transport and switching, it is highly attractive to extend IP over the air interface to the end-user equipment rather than terminating at the RAN. This eliminates the dependencies between applications and the wireless access network and expands the opportunity for more players to participate and develop new applications. This should be compared with present-day cellular services (see the section Current State of Cellular Systems), which are vertically integrated and optimized for voice, resulting in high performance in terms of spectral efficiency but low flexibility in introducing new services. The network architecture we are envisioning for the future is shown in Figure 2, where all links (wireline and wireless) will carry IP traffic.

CURRENT STATE OF THE INTERNET

The term "Internet" refers to a wide collection of data networks joined together through the common use of Internet protocol (IP) and its associated routing and addressing conventions. Each network in the collection can be thought of as a separate autonomous system (AS) that takes responsibility for delivering data packets within the system however it sees fit while conforming to a standard protocol at exchange points with other participating networks. Some examples of ASs include corporate networks and ISPs. This decentralized architecture is one of the most distinguishing characteristics of the Internet. In this section, we review some of the more salient features of the Internet.

Table 1 World Cellular Subscriber Forecasts

Millions of subscribers	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005
World	136	204	307	474	728.8	988.7	1,293.0	1,580.2	1,825.3	2,029.5

Source: EMC World Cellular Database, October 2001, based on actual figures to end of June 2001.

Table 2 World Internet Users

Millions of subscribers	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005
World	55	101	150	201	407	518.5	664.7	813.7	978.8	1156[a]

Source: Nua Ltd (up to year 2001) and eTForecasts (forecasts from 2002–2005).
[a]Extrapolated from estimates for 2004.

Classless Addressing

Each node or a device on the Internet must have a unique IP address. Messages are routed through the IP network using the IP address of the destination. The most widely deployed version of IP today is version 4 (IPv4) (Postel, 1981a), which uses 32 bits of address space. The IP address is typically written as four separate numbers, each coded using 8 bits, separated by periods (an example would be 128.200.222.100). In an isolated network, any IP address that matches the format may be used (except for certain reserved addresses), as long as each node or device has a unique address. However, to connect to the Internet, it is necessary to obtain addresses not being used so as to avoid having duplicate addresses with other networks and nodes.

In the early days, the InterNIC assigned to anyone who asked a range of addresses that belongs to one of three classes. Class A is the largest class, which supports up to 16 million hosts on each of 127 networks. A Class B network supports up to 65,000 hosts, and a Class C network supports up to 254 hosts. These classes were also used to route packets (see the section Dynamic Routing). Hence, this type of address allocation scheme has been termed classful addressing or routing. Although simple, this allocation scheme has proven to be very inefficient in terms of actual use of IP addresses. For example, a modestly sized network may need to support up to 3,000 nodes in its own network. Since a Class C network supports only 254 hosts, a Class B network would have been assigned even though less than 5% of the assigned addresses will be used. The possibility of running out of IP addresses became quite real as the growth of the Internet exploded. Although IPv6 (Deering & Hinden, 1998) can solve the address crunch by increasing the address space to 128 bits, it does not resolve the nearer term problems described next.

Since Class A networks are far too large for most organizations, and Class C networks are too small, Class B networks were the most commonly requested and granted type. In August 1990 during the Vancouver Internet Engineering Task Force (IETF) meeting, Frank Solensky, Phill Gross, and Sue Hares projected that the Class B space would be depleted by March 1994. This led the InterNIC to force many smaller organizations to accept several Class C networks rather than a single Class B network. In the example above, 3,000 nodes would require at least 12 Class C networks. The problem with this situation is that each network must be routed individually when classful routing is used, so instead of having a single entry for the network in the routing table, there would now be 12 entries in the table. This led to extremely large routing tables in backbone routers, slowing down the whole network.

Classless Interdomain Routing

To address this problem, classless interdomain routing (CIDR) (Hinden, 1993) and classless addressing (Rekhter & Li, 1993) were introduced in 1993. Instead of using one of just three partitions between the network and the host portion of the IP address (determined by the "class" of the network), it became possible to have variable length network identifiers or "prefixes." CIDR can have any prefix length between 13 and 27 bits instead of just 8, 16, or 24 bits in classful routing. Thus, the smallest network can have up to 32 hosts and the largest network can have more than 500,000 hosts. A CIDR address includes the standard 32-bit IP address as well as the information on how many bits are used as the network prefix. Increasing the efficiency of address allocation has reduced not only the address crunch, but also the size of the routing table. CIDR can also be used for route aggregation, which further reduces the size of the routing table. Deployment of CIDR has contributed greatly to the continued growth of the Internet.

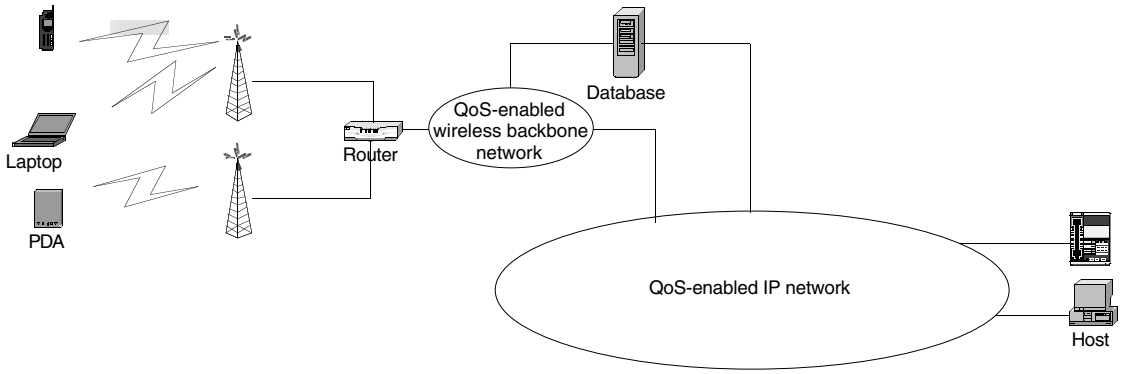


Figure 2: Architecture of packet-switched wireline/wireless multimedia network.

Dynamic Routing

In a circuit-switched connection as used in a public switched telephone network (PSTN), the path of a call is established at the beginning of a connection and is maintained throughout the duration of the call. One of its weaknesses is the vulnerability of the network when central switching stations fail. A much worse problem for data traffic has been the fact that such traffic tends to be bursty and the reserved circuit is unused much of the time, thus wasting much of its capacity. In order to increase the efficiency and robustness, a packet-switched architecture was envisioned for the Internet. Packets from many sources and destinations may share a common transmission circuit but be switched independently, and where the flow of data packets is constantly monitored and adjusted through the network, possibly around any failed nodes. This process of directing the switching and flow of packets is called "routing."

Routing in the Internet is performed by a device called a router, which has connections to multiple networks and has the ability to relay data packets between these networks. The router decides to which network each packet should be sent based on the information available in the IP header of each packet. Currently, most routers look at only the destination address to decide the fate of a packet, even though other information is available (see the section Real-Time Traffic Support). Instead of the entire IP address, only a subset of the address is used by the router. In classful routing, only the class portion of the address is used, while in CIDR, only the prefix portion of the address is used. Based on the destination address, the router consults its routing table, which lists the port number associated with the precomputed "optimal" path for that destination address. The routing table can be configured manually or automatically. In either case, the optimal path is determined based on the network topology and different link metrics (e.g., bandwidth, delay, load, reliability, and cost).

When the routing table is configured manually, static routing is said to be used. It is the preferred mode of operation in networks with few nodes or in stub networks, which is a network with only one or two paths to the rest of the network.. While simple to configure for small networks, a router using static routing cannot reroute packets automatically if the instantaneous traffic loads change or if a router or a link in a preconfigured route goes down for any reason. The destination may remain unreachable until human intervention is made to update the routers, based on new traffic loads or utilization of the failed link or node. Dynamic routing addresses these deficiencies by enabling dynamic and automatic update of the routing tables in routers. For example, if a router detects failure of a link, it will notify other routers of the condition so that appropriate adjustments to the routing table entries can be made. The algorithm for calculating the optimal path and the mechanisms for sharing information among different routers specify numerous routing protocols used in the Internet.

Routing protocols are generally divided into two groups. Interior gateway protocol (IGP) is used within ASs, and exterior gateway protocol (EGP) is used between routers of different ASs, i.e., across the Internet. The two most widely used IGPs are routing information protocol (RIP) (Hendrick, 1988) and open shortest path first (OSPF) (Moy, 1998b). Examples of EGP include somewhat confusingly termed exterior gateway protocol (EGP) (Mills, 1984) and border gateway protocol (BGP) (Rekhter & Li, 1994). For details about the protocols, interested readers may consult the RFCs as well as numerous books and articles (Doyle, 1998; Huitema, 2000; Moy, 1998a). Below we present a brief overview of the most popular routing algorithms currently in use in the Internet.

Routing Information Protocol (RIP)

RIP is a distance vector protocol that uses the number of "hops" (i.e., the link between two nodes) to determine the optimal path to the destination. Every 30 s, the entire routing table is broadcast, and the routing tables of the listening routers are updated based on the reported hop counts. This reliance on second-hand information results in relatively low computation and storage requirements, but it also results in slow convergence. RIP is suitable only for relatively small networks (no wider than 15 hops) since the maximum hop count is restricted to 15 (a hop count of 16 is considered infinity). In addition, the high bandwidth overhead of broadcasting entire routing tables makes the algorithm hard to scale.

Open Shortest Path First (OSPF)

OSPF is a link state-based algorithm, in which the link states of each node are flooded to all routers in the network. The transmission takes place only when there is a change in network topology. The link state may convey information about various link states such as throughput, delay, loss rate, or some cost function. Based on the link states, each node in the network computes the complete network topology. As all nodes have access to the same link state information, all nodes should arrive at the same network topology. Based on this map of the network, the shortest path tree to each destination is computed using a shortest path first algorithm, and the result is used to populate the routing table. Such computation places much heavier burden on memory and processing power than algorithms such as RIP, but it also results in faster response to network events such as a link failure, quicker route convergence, and requires less traffic overhead especially for larger ASs. OSPF also offers additional advanced features that are described in detail in the literature. Such advantages are driving the increasing use of OSPF over RIP.

Border Gateway Protocol (BGP)

BGP is an interautonomous system (inter-AS) routing protocol. BGP version 4 has become the main inter-AS routing protocol in the Internet. BGP is a path vector protocol, which defines a route as a pairing between a destination and the attributes of the path to that destination. A BGP router learns of multiple paths via internal and external BGP speakers and picks the best path, which is then sent to external BGP neighbors. A network administrator has control over the policies applied during the best path selection. In addition, unlike other routing protocols,

BGP is connection oriented and uses TCP as the transport protocol. BGP supports IP prefixes and path aggregation, which makes it suitable for CIDR.

Real-Time Traffic Support

An important class of applications used over the Internet is real-time. Real-time applications generate traffic that must be communicated to the recipient(s) within a very short amount of time. The time limit may be milliseconds for interactive applications such as voice and video telephony, but it may be larger for certain streaming or transaction processing applications. Despite increasing interest in supporting such applications over the Internet, the Internet is still largely a best-effort network without any delivery guarantees necessary to provide the levels of quality of service (QoS) expected from end-users.

QoS Support

Different applications typically have different traffic characteristics and QoS requirements. Applications such as streaming video usually require large bandwidth, and interactive applications such as voice over IP (VoIP) and video over IP (VIP) require tight delay bounds as the data must be played back continuously at the rate they are sampled. If the data (packet) does not arrive in time, the playback process at the receiver will be disturbed. For example, in voice telephony, human beings can tolerate a latency of up to about 200 ms (Brady, 1971), although in most of today's voice networks, the latency is limited to around 50 ms. If the latency exceeds this value, the degradation in call quality will be noticeable. If enough extra bandwidth is available, best-effort service may be able to fulfill the delay, throughput, and other requirements. When resources are scarce, however, real-time traffic will suffer from congestion and delay, resulting in degradation in the application quality.

DiffServ. To facilitate end-to-end QoS on IP networks, the IETF has defined two models: Integrated Services (IntServ) (Braden, Clark, & Shenker, 1994) and Differentiated Services (DiffServ) (Blake et al., 1998). IntServ was defined first and follows the signaled-QoS model, where the end-hosts signal their QoS needs to the network using resource reservation protocol (RSVP) (Zhang, Deering, Estrin, Shenker, & Zappala, 1993). RSVP signaling and reservation of the desired QoS is done for each flow in the network. A flow or a stream is defined as an individual, unidirectional data stream between two applications, and is uniquely identified by a 5-tuple (source IP address, source port number, destination IP address, destination port number, and the transport protocol). While IntServ provides for a rich end-to-end QoS solution, there are several problems with the approach. State information for each reservation needs to be maintained at every router along the path, and each packet must be processed to ensure that the QoS of each flow is being satisfied. As there may be hundreds of thousands of simultaneous flows going though a network core (router), it is not clear whether IntServ will scale well in terms of complexity of admission control, memory requirements of maintaining state, and packet processing overhead.

Since per-flow QoS is difficult to achieve in an end-to-end fashion without introducing scalability issues, it naturally leads one to think about classifying flows into aggregates (classes), and providing QoS to aggregates rather than to individual flows. For example, all real-time flows could be grouped into a single class, and bandwidth and other resources can be allocated for the class. This would reduce the router's burden on classification of traffic, signaling, and state maintenance requirements. This is the approach taken in the DiffServ model. In this model, packets are first divided into classes by marking the type of service (ToS) byte in the IP header. A 6-bit bit-pattern called the Differentiated Services Code Point (DSCP) in the IPv4 ToS Octet or the IPv6 Traffic Class Octet is used to this end. Once packets are classified at the edge of the network, specific forwarding treatments, formally called per-hop behavior (PHB), are applied at each network element, providing the packet with appropriate guarantees (deterministic or statistical) on performance metrics such as delay, delay jitter, and bandwidth. This combination of packet marking and utilization of PHBs results in a more scalable QoS solution.

MPLS. Multiprotocol label switching (MPLS) (Rosen, Viswanathan, & Callon, 2001) is another emerging technology that seeks to introduce QoS guarantees on the Internet. Existing link state protocols, specifically OSPF and IS-IS, provide the link state information about the underlying IP network. Such information is used to determine the best path through the network called label switched paths (LSPs) using criteria such as number of hops and other configurable parameters such as delay and bandwidth.

An incoming packet to an MPLS network is assigned a "label" by an "edge-label switch router" (Edge-LSR). This label is swapped by intermediate label switch routers (LSRs) as the packet traverses the MPLS network on an LSP, and the final label is removed when leaving the MPLS network. The label distribution protocol (LDP) is used to establish label-to-destination network mappings. Forwarding of a packet is based solely on the contents of the label, and not on the IP headers as is done in normal IP routing, speeding up the process. Such increase in performance, as well as the ability to perform traffic engineering, makes MPLS a strong contender for the converged network.

Real-time Transport Protocol (RTP)

The real-time transport protocol (RTP) (Schulzrinne, Casner, Frederick, & Jacobson, 1996) is an IP-based protocol providing support for the transport of real-time data traffic such as video and audio streams. The services provided by RTP include timestamping, sequence numbering, and other mechanisms to take care of the timing issues. RTP also provides information about the encoding method used in the underlying data. Through these mechanisms, RTP provides end-to-end transport for real-time data over the IP network. RTP was primarily designed for multicast of real-time data; unicast is supported as well. It also can be used for one-way transport such as video-on-demand service as well as interactive services such as VoIP and VIP.

RTP was designed to work in conjunction with an auxiliary control protocol called real-time control protocol (RTCP) (Schulzrinne et al., 1996). In an RTP session, participants periodically send RTCP packets to convey feedback on quality of data delivery and information of membership. RFC 1889 defines five RTCP packet types to carry control information. These packets contain information regarding number of packets lost, interarrival jitter, and timestamps. Through these control information packets, RTCP provides services such as QoS monitoring, congestion control, intermedia synchronization, and calculation of round-trip delays. RTP is typically run on top of user datagram protocol (UDP) (Postel, 1980) to make use of its multiplexing and checksum functions. However, use of these protocols introduces bandwidth overhead to each data packet, which is especially important in low-speed wireless links.

Transmission Control Protocol (TCP) and UDP

Transmission control protocol (TCP) (Postel, 1981b) is the most commonly used transport protocol on the Internet. TCP provides a connection-oriented and reliable flow between two hosts, while UDP provides a connectionless and unreliable datagram service over the network. UDP was chosen as the target transport protocol for RTP because of three reasons. First, since RTP was initially designed for multicast, it was realized that connection-oriented TCP does not scale well for a large number of flows and therefore is not suitable. Second, for real-time data, 100% reliability is not as important as timely delivery. Since TCP provides reliability through retransmissions, it is not suitable for real-time applications, for by the time packet error or loss is detected and the retransmitted packet is received, the playback time of the data contained in the packet would likely have passed. Thus, retransmission only increases the network traffic without benefiting the quality of the playback. Third, congestion control of TCP does not match well with the needs of real-time applications, as real-time applications cannot tolerate packets being held back even during periods of congestions.

Internet Applications: World Wide Web, E-mail, Instant Messaging

Since the advent of the Internet, a handful of applications have become the drivers of the growth of the Internet. Some of these applications include e-mail, file transfer, instant messaging, and the World Wide Web (WWW). File transfer was one of the original applications envisioned for the Internet and continues to be one of the most popular applications. Typically, file transfer protocol (FTP) (Postel & Reynolds, 1985) is used to send and retrieve files from a remote computer. E-mail was initially used to send text messages between individuals or groups of individuals, but increasingly diverse media such as images and audio are being transmitted using e-mail. According to the Nielsen/NetRatings First Quarter 2002 Global Internet Trends report (2002), e-mail was the most dominant online activity in 12 countries over the previous six months.

Three technologies were invented by 1991 to accommodate the arrival of the WWW. Hypertext markup language (HTML) (Raggett, Le Hors, & Jacobs, 1999) is used to produce Web documents, hypertext transfer protocol (HTTP) (Fielding, Gettys, Mogul, Frystyk, & Berners-Lee, 1997) is used to transport HTML documents from the server to the client, and the client Web browser is used to retrieve, interpret, and display HTML documents. Perhaps no application has been more instrumental in piquing the general public's interest in the possibilities of the Internet. In March of 1993, WWW traffic measured mere 0.1% of NSF backbone traffic. By February 1995, WWW passed FTP as the largest volume Internet application. The main attraction of WWW lies in its flexibility to service documents containing various media including text, graphics, audio, and video, as well as the ease of accessing documents using "hotspots."

Instant messaging (IM) is an application that has been experiencing a tremendous growth of late. IM enables individuals to create private "chat" sessions, in which various types of messages may be exchanged. It offers all the capabilities of another immensely popular application, e-mail, but with near real-time response. The growth of IM has been strong for both home use and business use. A November 2001 study from Jupiter Media Metrix showed that the number of unique business users of the top three IM applications (AOL, MSN, and Yahoo!) increased from 10 million in September 2000 to 13.4 million in September 2001. During the same time period, the total usage time increased from 2.3 billion minutes per month to 4.9 billion minutes per month. The growth in home use has been equally impressive with the number of users reaching 53.8 million and the total usage time reaching 13.6 billion minutes a month. IM service is very similar to short messaging service (SMS) available in the wireless networks, and the increasing use of IM over the Internet will only increase the demand for SMS (Instant messaging, n.d.).

CURRENT STATE OF CELLULAR SYSTEMS (FOCUS ON 2G)

Cellular Layouts

Most wide area wireless networks today are cellular. The service area is divided into smaller service areas called cells. In contrast, the first mobile telephony systems in operation up to late 1970s did not use the cellular layout. Instead, they relied on a high-power transmitter to service a large area. Because only a fixed number of frequency channels were assigned to each service area, increasing demand for mobile telephony meant increasing competition for the available channels, resulting in excessively high call blocking rates. The cellular system was designed to address this issue, to increase capacity through the employment of frequency reuse. The first systems that used cells were based on analog technology and are referred to as first generation (1G) systems. In some places, such as the United States, the analog network still exists to serve customers that have not yet transitioned to the next generation of cellular systems that use digital transmission technology. The digital cellular service is referred to as the second generation (2G) system. By going over to

digital transmission, the cellular systems gained tremendous flexibility because of the inherent intelligence that could now be built into the management and control of the system. In addition, digital signals can be pre- and post-processed to enhance the received signal quality. Third generation systems, besides being all digital, promise to offer higher bit rates for data services and to switch from circuit- to packet-based transmission. To address the need for higher bit rates for data services, some intermediate technologies that fit into the 2G architecture were defined. We discuss those in the section Higher Bit Rates for Data—GPRS and HDR. A more detailed description of 3G is given in the section 3G Cellular Systems and in the final section of the chapter we very briefly discuss the directions for 4G.

The advantage of creating cells is that after dividing the service area into cells, the same frequency sets can be used and reused systematically. Because the same set of frequency bands can be reused many times, a larger number of users could be supported (Jakes, 1993; Yacoub, 1993). Each cell has a base station, which contains antenna and radio equipment, as well as a high-speed, high-capacity connection to the network. Since the area of the cell is typically much smaller than the carrier's coverage area, lower transmit power can be used to communicate with the mobile station (MS). However, even with lower transmit power, signals still propagate into neighboring cells, causing interference. Thus, cells with the same frequency sets are spaced many cells apart, and immediately neighboring cells use different sets of frequency channels to reduce interference (one exception is CDMA systems, such as IS-95, in which the same frequencies can be used in all cells due to its ability to work at low carrier-to-interference ratios) (Viterbi, 1995). A group of nearest cells that use disjoint sets of frequency bands is called a cluster. The service area is composed of these clusters, which reuse all frequency bands allocated to the service provider.

Most current cellular networks utilize an additional level of subdivision to further increase capacity beyond that achieved using the cellular concept. Instead of making cells even smaller (which introduces the problem of frequent handoffs, which will be discussed later), each cell is divided into 3 or more sectors (Yacoub, 1993). Previously, an omnidirectional antenna was used at the base station to transmit signal equally in all directions, which also distributes interference equally in all directions. The omnidirectional antenna is replaced by several directional antennas, each of which can direct the radio wave to a certain direction. For example, an omnidirectional antenna may be replaced by three directional antennas, each with a beam width of 120°. Thus, the cell is effectively sectorized into three distinct areas, each with its own set of frequencies. Since the signal is transmitted only in the sector that contains the MS, sectorization reduces the interference between cells in adjacent clusters. Reduced interference means that clusters can be located closer together, enabling more frequent reuse of the same frequencies and subsequently increasing traffic capacity.

A major obstacle to the cellular concept working effectively involved a mobile user traveling from the coverage area of one cell to that of another during a call. This problem did not exist, at least not to the same extent, in noncellular systems where the service areas were much larger. However, as the sizes of cells shrank (in order to maximize frequency reuse), it became very likely for a user to travel outside the coverage of a cell, and simply dropping those calls was unacceptable from the point of view of the mobile user. To address this concern, handoff techniques were developed as a part of different wireless standards. By using handoffs, it became possible to automatically transfer a call from a radio channel in one cell to that in another without disrupting the ongoing connection. This ability made the adaptation of the cellular concept practical and realizable.

Mobility and Roaming

One of the primary benefits of wireless telephony is the ability to move around without the concern of losing connectivity, at least most of the time in most areas. However, typical mobile calling plans specify a home calling area, such as a particular metropolitan area, a state, or even the entire United States. When a mobile user travels outside this area, he is said to be "roaming." Even if the home calling area covers a large area, it is possible and quite likely that in certain geographical locations, the user's carrier does not have coverage while another carrier may. Thus, roaming capability is necessary in order to increase the level of any time, anywhere access, to which mobile users have become accustomed and expect. The latter type of roaming is made possible through the business agreements between carriers and service providers to grant each other's customers access to their networks, in addition to technologies and standards summarized below.

Two of the most popular technologies enabling the roaming service are GSM Mobile Application Part (MAP) (MAP Specification, 1997) and IS-41 (Telecommunications Industries Association, 1991). Both have become standards adopted in their respective areas of application. MAP is used in GSM networks (Mouly & Pautet, 1992), while IS-41 is used in IS-136 (Telecommunications Industries Association, 1996) and IS-95 networks (Telecommunications Industries Association, 1993). IS-41 was developed a few years after the development of GSM MAP, and adopted terminology, network architecture, and some protocol details from MAP. Three major components are used to enable the roaming service: the home location register (HLR), the visitor location register (VLR), and the mobile switching center (MSC). The HLR is the database that contains information about subscribers in the network, including the current locations of the subscribers. For roaming subscribers, the location is in the form of the signaling address of the VLR associated with the subscriber. The VLR temporarily stores a subset of information available in the HLR for those mobile users whose HLRs are located elsewhere. The MSC is a telephone exchange that is able to set up and route mobile calls. For each user in its service area, the MSC utilizes either the HLR or the VLR to setup and route calls. In both IS-41 and GSM MAP, common channel signaling system number 7 (SS7) (Black, 1997) is used to exchange call setup and routing information, including HLR and VLR access, over a digital signaling network.

Voice Telephony as the Primary Service

The current generation of the wide area wireless networks, i.e., the cellular networks, has been developed primarily to serve voice telephony. From the physical layer (e.g., channel coding and transmission) to the application layer (e.g., voice compression), all aspects of the system were designed and optimized for the purpose of maximizing the capacity and the quality of phone calls. Despite the increasing use of services such as wireless Web and SMS (to be discussed later), the bulk of present-day wireless traffic remains voice telephony. For example, according to Strand Consult's report from 2002, more than 85% of mobile service revenues in Europe were from voice telephony in year 2002 (How to make money, 2002). This is especially noteworthy since Western Europe is one of the leading markets in use of non-voice wireless services.

Another leading market is South Korea. SK Telecom, Korea's leading mobile carrier with 11 million subscribers which claims to have the most advanced mobile network based on CDMA 2000 1X (Telecommunications Industries Association, 2001), recently reported that it expects to gather only 5.4% of its expected total sales from mobile Internet services. Even though this figure represents a large increase over previous years, it remains small in comparison to the revenue from voice telephony. The lesson here is that while the growth of nonvoice services over cellular networks will continue, it is important to continue to serve what is currently the most valuable market, i.e., voice telephony.

Popularity of Instant Messaging (SMS)

Since its launch in 1995 as a part of the original GSM specifications, short messaging service has become a tremendously popular service offered on wireless telephone networks. Although its nascent growth was fueled by the younger generation, by the year 2000 the popularity of SMS has spread beyond the original group. Worldwide, the use of SMS has skyrocketed. The GSM Association announced in February 2001 that 15 billion messages were sent over the world's GSM wireless networks during December 2000, compared to only 3 billion messages a year before. In terms of revenue, research outfit IDC expects SMS revenue to reach US$6.5 billion worldwide by 2002.

SMS provides the ability to send and receive text messages to and from mobile handsets with message lengths ranging from 120 to 256 characters. The communication is near real-time as in Internet-based IM. In addition to GSM, SMS or SMS-like services are available in other popular wireless standards, such as in IS-136 and IS-95 networks. Virtually any type of information based on text can be sent using SMS, including e-mails, news headlines, and some games. The range of applications of SMS has yet to be thoroughly explored and continues to expand.

The tremendous success of SMS has resulted in development of more advanced versions of SMS. The enhanced messaging service (EMS) is an open 3rd Generation Partnership Project (3GPP) standard (3GPP TS 23.040, 2002) that allows a mobile phone to send and receive not only plain text, but text enhanced with different fonts and sizes, images, sounds, and animation. Since EMS also utilizes the signaling channels for transport and the same SMS Centers, no network modifications are needed to support SMS, which allows a relatively painless upgrade path.

In comparison, multimedia messaging service (MMS) offers more drastic changes both in terms of functionalities and requirements (3GPP TS 23.140, 2002). For example, MMS data are delivered over the traffic channel rather than the signaling channel. This would require one of the new mobile network infrastructures, such as General Packet Radio Service (GPRS) (Cai & Goodman, 1997) or 3G, as well as new network elements such as multimedia messages relays and servers to fully utilize the capabilities of MMS. In return, rich media messages composed of text, images, audio, and video will be made possible. Some of the envisioned applications include maps, cartoons, games, and interactive videos. Just as the current success of SMS was totally unexpected, it is imprudent to speculate on the future of these newfangled technologies. With the development of suitable applications, however, these services have a good chance of success.

Web Access

Based on the growth from the traditional methods of Web access based on modems and other wireline devices, wireless Web access was expected to be the next killer application on the wireless cellular networks. Who wouldn't want to access all those Web sites with the conveniences of not being tethered to any sockets or switches on walls? However, the lofty expectations of exponential growth have yet to be realized. Problems such as limited speed, incompatible data standards, and poor user interfaces have hindered the growth. Even so, there were already more than 28 million subscribers in Japan, or close to 1/5 of the population, who subscribe to NTT DoCoMo's i-mode mobile Internet service at the end of April 2002 (technically, i-mode is a specification, not a standard).

Fortunately, there has been much ongoing progress in addressing some of the difficulties that have plagued wireless Web access. Higher bandwidth technologies such as GPRS and high data rate (HDR) CDMA (Bender et al., 2000) are beginning to be deployed and will increase data transmission speeds to more than 171 Kbps. Advances in user interfaces such as voice recognition and better displays, in conjunction with increasing use of industry standards such as WAP and i-mode, will continue to improve accessibility of the Web via the wireless network.

Higher Bit Rates for Data—GPRS and HDR

Two of the more promising wireless technologies slated to become the main drivers of wireless data access are GPRS and HDR. Whether they are true 3G service or mere stopgap 2.5G is up for debate, but both are able to provide significantly enhanced data rates, which will allow more diverse applications to be served over the wireless network.

GPRS

GPRS builds upon the tremendous success of GSM technology, which, according to the GSM Association, claims more than 825 million users in over 193 countries as of March 2003. The major advantage of GPRS over the existing GSM data services such as circuit switched data (CSD)

and SMS include much enhanced speed, "always-on" connection, and spectrum efficiency.

Although the underlying GSM networks uses a circuit-switched architecture, GPRS uses packet-switching technology to achieve "always-on" connectivity and higher spectrum efficiency. All eight timeslots in a frame can be used in GPRS to achieve the peak rate of 171.2 Kbps. Unlike in CSD, the timeslots are not set aside to a user for the duration of the connection. Instead, radio resources are used only when data are actually being transmitted. Thus, the same channel can be shared among many users concurrently. This efficient sharing of the limited bandwidth allows the network operator to maximize use of the limited radio resources, and lower the cost for the mobile users.

Since the existing GSM MSCs are based on circuit-switching technology, they cannot handle the operation of packet-switched GPRS connections. Thus, two new network components called serving GPRS support node (SGSN) and gateway GPRS support node (GGSN) have been added to the GSM architecture. The SGSN behaves much like the MSC, but for GPRS traffic. It is responsible for delivering packets to the mobile users in its service area, and also handles queries to HLR for roaming operation. The GGSN is the interface to the external networks, such as the Internet. It maintains address and routing information necessary to tunnel data packets to appropriate SGSNs and onto the MSs. This makes it possible to use existing IP applications over the GPRS network. This and other advantages of GPRS already mentioned may help initiate the long-awaited proliferation of wireless Web access.

HDR

HDR is based on CDMA technology and achieves a peak data rate of 2.4 Mbps. Instead of sharing the channel with voice data as is done in GPRS, in HDR the entire frequency channel is allocated to data traffic. By decoupling the data service from voice service, the overhead required to meet strict latency requirements of voice calls no longer degrades the system's ability to handle packet data efficiently.

Large improvement in spectrum efficiency is achieved by measuring the signal-to-noise-plus-interference ratio between the BS and the MS, and adapting the modulation scheme and the forward error correction to achieve the optimal efficiency for the given channel condition. Since only a single user is served at any given time (in units of 1.67-ms packets), there is no degradation in capacity due to inference among the MSs. In addition, by taking advantage of varying and more relaxed latency requirements of data traffic, further increase in throughput can be achieved (Bender et al., 2000).

The network architecture of HDR has been designed with the Internet in mind. Selection of the point-to-point protocol (PPP) (Simpson, 1994) and the PPP multilink protocol (MP) (Sklower, Lloyd, McGregor, Carr, & Coradetti, 1996)were based on the need to support IP traffic with different QoS requirements while utilizing low overhead. In addition, the radio link protocol (RLP) has been designed to achieve the level of data fidelity, i.e., bit error rate, which PPP and IP experience in wireline networks. This is important since many upper layer protocols, such as transmission control protocol (TCP), had been designed and optimized for the conditions observed in wireline networks. Such network architecture and enhanced data rates are expected to improve and enhance the usability of wireless data services.

WIRELESS INTERNET—IS IT HAPPENING?

Mobile IP (MIP)

Mobility is not a feature that was incorporated into IP when it was conceived several decades ago. Mobile IP consists of the necessary extensions needed to support mobility in the Internet (Perkins, 1997). IPv4, currently the most prevalent IP version on the Internet, has no provisions for mobility. Each host computer on the Internet has a unique address in the hierarchical IP addressing space. Each address consists of a network prefix and a host number. The network prefix of the address determines the location (i.e., campus network) of the host computer. Routers do not contain the address of each individual host in their routing tables; instead, the network prefix of each address is used to forward packets to the next hop en route to the destination network. To reach a host, you must know its IP address; if a host moves, all traffic addressed to it will be sent to its home location following the rules of network prefix routing (see Dynamic Routing for more details). If a host is assigned an address at its current new location, the sender must be made aware of it so that the appropriate destination address can be used in the data packets. As the host moves, its IP address will change to reflect its new location. This means that the sender must be informed of every change in location/address to maintain the connection and data flow. Packets in transit will be lost unless some provision is made to have them follow the host to the next location. For TCP connections the problem is further exacerbated as the IP address of the host is used in the TCP connection for session identification purposes. Thus, if the host moves and changes its IP address, the connection identifier will no longer be valid, thereby causing the session to be terminated. It is for these reasons that MIP was conceived. For true mobility, the whereabouts and mobility of a host should not affect its ability to be reached by any sender, and in addition, the sender should not have to be responsible for tracking a remote host as it moves about the Internet.

When designing MIP, it was obvious that it had to be

Compatible with the existing installed base of IP and the layers below it;

Transparent to the layers above IP;

Scalable and efficient, and capable of supporting large numbers of hosts and not impede the functionality of the Internet in anyway; and

Secure, and the forwarding of connection control information must be authenticated to prevent traffic from being diverted to other destinations.

The design of MIPv4 accomplished the above four requirements. By all means, it is not the most optimal solution.

Fortunately, IPv6 was designed with flexibility and mobility in mind and, as such, can support mobility in a more optimal/natural manner. Unfortunately, IPv6 is not being adopted as quickly as expected, and is currently only being used in some small isolated locations and uses IPv4 tunnels for connectivity.

Below we will highlight some of the major features of MIPv4.

MIP Terminology

In RFC 2002 describing IP mobility support (Perkins, 1996), the following entities were identified in conjunction with MIP:

Mobile node (MN)—A device that supports MIP and can change its location without affecting its communication abilities so long as layer 2 connectivity is available.

Home agent (HA)—A device in the home network (i.e., the subnet to which the MN's IP address belongs) of the MN that keeps track of the location of an MN. It tunnels packets destined for the MN to its new address. The HA is typically a router on the home network.

Foreign agent (FA)—A device in the current foreign network (i.e., the network that the MN is currently visiting) of the MN that can forward packets sent to it for the MN to the MN if it terminates the tunnel set up by the HA. The FA is typically the default router on the foreign network.

Care of address (COA)—IP address that defines the current location of the MN. It is the address to which the HA forwards all packets for the MN, thus terminating the tunnel. There are two possible scenarios for COA:

1. The COA is located at the FA; i.e., it is the IP address of the FA. The FA then terminates the tunnel and forwards the packets to the MN. This approach allows many MN to share one IP address.
2. The MN temporarily acquires a new IP address. For this scenario, the MN terminates the tunnel. Although a convenient approach, it does require that several IP addresses be made available for mobile devices, which may not always be the case. This is referred to as a colocated COA.

Correspondent node (CN)—A device that communicates with the MN. It is unaware of the location of the MN and just simply uses the MN's original IP address for packet forwarding.

From the above we see that a tunnel starts at a HA and terminates at either the FA or the MN. The HA keeps track of where the MN is. The FA is not always needed for MIP functionality although it may be necessary for security purposes.

Operation of MIP

The operation of MIP consists of three steps: agent discovery, registration, and encapsulation/routing/tunneling. Below, we discuss each step:

Agent discovery—Consists of broadcast messages used by the MNs to detect whether they have moved. These messages are sent out periodically by the default router (FA) on a subnet. If the MN has not heard an advertisement, it will solicit for one.

Registration—All MNs are required to register with the HA and the FA (if used). As registrations expire, MNs must re-register periodically. Any move to a new location requires a new registration.

Encapsulation/routing/tunneling—All packets that arrive for the MN on the home subnet are claimed by the HA. The HA proceeds to encapsulate them to reflect the new COA of the MN and then routes all the traffic to the MN on the foreign network. The CN sends all messages to the MN's IP address, and the HA relays them, via the tunnel, to the COA. The MN sends its messages directly to the CN. This type of communication results in what is referred to as triangular routing: CN to HA, HA to MN, MN to CN. To improve the performance of MIP, route optimization has been proposed, which allows the CN to learn the COA of the MN and correspond with it directly. This does mean that the CN must be informed of any change in location of the MN and requires additional authentication procedures.

Security in MIP

Security is one of the main concerns in any mobile environment. It is necessary that all devices involved with data reception and forwarding be authenticated to ensure their identities. MNs must register periodically with the HA, which involves an authentication process. If an MN moves, it must re-register at the new location. If route optimization is used, the CN must authenticate itself before being capable of communicating directly with the MN.

TCP for Wireless Networks

The TCP protocol was not designed to operate over channels that are lossy in nature. It uses timers at the sender to determine the state of congestion in the network. If an ACK does not come back before the timer expiration, it assumes that the link is congested and shuts itself down by decreasing the congestion window size to one. It then proceeds to retransmit the unacknowledged packet. TCP does not take into account the possibility of a lost packet due to channel conditions as a lost packet is directly interpreted as congestion. Because of the proliferation of wireless networks, it is imperative that the transport layer understands the difference in performance of the wireless link and not make erroneous assumptions as to the state of the network. Several papers that propose modifications to TCP to improve its performance over the wireless channels have appeared. We describe some of these approaches below:

Indirect TCP (I-TCP) (Bakre & Badrinath, 1995)—It segments the TCP connection into two portions: one for the wireline transmission, and one for the wireless segment. It uses the traditional TCP over the wireline segment, and a modified version for the wireless segment. Assuming the use of MIP, the FA is the most likely candidate for acting as the proxy and acknowledging all packets as well as terminating the connection. Over the wireless segment, the proxy communicates with the

MN using local ACKs not relayed to the CN. This mechanism shortens the retransmit time (most retransmissions occur only over the short wireless portion). However, should the FA fail, packets will be lost and neither side of the connection will be aware that this happened. This scheme violates the end-to-end semantics of TCP.

Snooping TCP (Balakrishnan, Seshan, & Katz, 1995; Brewer et al., 1998)—This variation on TCP does not violate the end-to-end semantics of TCP as I-TCP does. It resides on an intermediate node (such as a FA in the case of MIP) that is attached to the wireless link. It buffers all data that are meant to be transmitted over the wireless channel so that if a loss occurs, it can retransmit the buffered data immediately, thereby avoiding the long end-to-end delay that otherwise would occur and cause the session to time-out. Although it does not violate the end-to-end semantics, it is not as efficient as I-TCP as it does not completely hide the wireless link from TCP; the retransmissions do incur a delay that could impact the performance of the end-to-end TCP session.

Mobile TCP (M-TCP) (Brown, & Singh, 1997)—It is similar in operation to I-TCP. However, no data are buffered or retransmitted by the proxy. The proxy only monitors the link, and if it determines that a loss has occurred, it will freeze the connection so that TCP does not go into slow start. On the wireline side it will close down the window, forcing the sender to go into persistent mode, and on the wireless side it uses a fast recovery TCP that does not use slow start. The end-to-end semantics are not violated as the proxy does not retransmit any data; it only interferes in the transmission by forcing certain TCP behavior on the detection of packet loss.

Other techniques include selective TCP (SACK) (Mathis, Mahdavi, Floyd, & Romanow (1996), which requires only the lost packets to be transmitted, and fast recovery/fast transmit (Caceres & Iftode, 1995), which does not allow TCP to go into slow start by sending three duplicate acknowledgments. All of the solutions above provide some features to prevent TCP from entering the slow start mode due to packet loss on the wireless channel.

IEEE 802.11b/g/a

Wireless LANs (WLANs) constitute an alternative technology to wireline LANs. They use radio signals to transmit data. Wireless technologies offer exceptional flexibility and mobility, thus making WLANs very attractive for many environments, including office, home, and public places. IEEE 802.11 (Institute of Electrical and Electronics Engineers, 1999)-type technologies are the common standard used in WLANs. Like all IEEE 802 standards, the 802.11 standards focuses on the lower two layers of the International Standards Organization (ISO) reference model, the physical layer and the data link layer.

Operation of 802.11

802.11 Operating Modes. 802.11 defines two pieces of equipment, a wireless station, which is usually a PC equipped with a wireless network interface card (NIC), and an access point (AP), which acts as a bridge between the wireless and the wireline networks. An access point usually consists of a radio, a wireline network interface (e.g., 802.3), and bridging software conforming to the 802.1d bridging standard. The access point acts as the BS for the wireless network, aggregating access for multiple wireless stations onto the wireline network. Wireless stations can be 802.11 PC Card, PCI, or ISA NICs, or embedded solutions in non-PC clients (such as an 802.11-based telephone handset).

The 802.11 standard defines two modes: *infrastructure* mode and *ad hoc* mode. In infrastructure mode, the wireless network consists of at least one access point connected to the wireline network infrastructure and a set of wireless stations. This configuration is called a basic service set (BSS). An extended service set (ESS) is a set of two or more BSSs forming a single subnetwork. Since most corporate WLANs require access to the wireline LAN for services (e.g., file servers, printers, and Internet links) they will operate in infrastructure mode. Ad hoc mode (also called peer-to-peer mode or an independent basic service set, or IBSS) is simply a set of 802.11 wireless stations that communicate directly with one another without using an access point or any connection to a wireline network. This mode is useful for quickly and easily setting up a wireless network where a wireless infrastructure does not exist or is not required or where access to the wireline network is barred.

802.11 Physical Layer. The three physical layers originally defined in 802.11 included two spread-spectrum radio techniques and a diffuse infrared specification. The radio-based standards operate within the 2.4-GHz industrial, scientific and medical (ISM) band. These frequency bands are recognized by international regulatory agencies, such as the FCC (USA), the ETSI (Europe), and the MKK (Japan) for unlicensed radio operations. Spread-spectrum techniques, in addition to satisfying regulatory requirements, increase reliability, boost throughput, and allow many unrelated products to share the spectrum without explicit cooperation and with minimal interference. The original 802.11 wireless standard defines data rates of 1 and 2 Mbps via radio waves using frequency-hopping spread spectrum (FHSS) or direct sequence spread spectrum (DSSS).

802.11b Enhancements to the PHY Layer. The key contribution of the 802.11b addition to the wireless LAN standard was to standardize the physical layer support of two new access speeds, i.e., 5.5 and 11 Mbps. To accomplish this, DSSS was selected as the sole physical layer technique for the standard since frequency hopping cannot support the higher speeds without violating current FCC regulations. The implication is that 802.11b systems will interoperate with 1- and 2-Mbps 802.11 DSSS systems, but will not work with 1- and 2-Mbps 802.11 FHSS systems.

802.11 Data Link Layer. The data link layer within 802.11 consists of two sublayers: logical link control (LLC) and media access control (MAC). 802.11 uses the same 802.2 LLC and 48-bit addressing as other 802

LANs, allowing for very simple bridging from wireless to IEEE wireline networks, but the MAC is unique to WLANs.

For 802.3 ethernet LANs, the carrier sense multiple access with collision detection (CSMA/CD) protocol regulates how ethernet stations establish access to the wire and how they detect and handle collisions that occur when two or more devices try to simultaneously communicate over the physical medium. In an 802.11 WLAN, collision detection is not possible due to the near/far problem. To detect a collision, a station must be able to transmit and listen at the same time, but in radio systems the transmission (near signal) drowns out the ability of the station to detect a collision (far signal).

Since collision detection is not possible, the stations only use collision avoidance; they sense the channel before transmitting. If the channel is busy, stations back off and try again at a later time. If the channel is idle, a station will transmit its frame. Since several stations may be sensing the channel at the same time and all detect it to be idle, they will start to transmit concurrently, thereby causing collisions. Because stations are unable to detect collisions, CSMA/CA systems need to use explicit packet acknowledgments (ACK). In other words, an ACK packet is sent by the receiving station to confirm that the data packet arrived intact.

Another MAC-layer problem specific to wireless is the hidden-terminal issue, in which two stations can both hear activity from the access point, but not from each other, usually due to distance or a physical obstruction. To solve this problem, 802.11 specifies an optional request to send/clear to send (RTS/CTS) protocol at the MAC layer. When this feature is in use, a sending station transmits an RTS and waits for the access point to reply with a CTS. Since all stations in the network can hear the access point, the CTS causes them to delay any intended transmissions, allowing the sending station to transmit and receive a packet acknowledgment without any chance of collision. Since RTS/CTS adds additional overhead to the network by temporarily reserving the medium, it is typically used only on the largest-sized packets, for which retransmission would be expensive from a bandwidth standpoint.

Security. 802.11 provides for MAC layer (OSI Layer 2) access control and encryption mechanisms, which are jointly known as wired equivalent privacy (WEP), with the objective of providing WLANs with security equivalent to their wireline counterparts. For the access control, the ESSID (also known as a WLAN service area ID) is programmed into each access point. A wireless client must know the ESSID to associate with an access point. No communication can occur unless there is an association between a client and the access point. In addition, there is provision for a table of MAC addresses called an *access control list* to be included in the access point, restricting access to only those clients whose MAC addresses are on the list.

For data encryption, the standard provides for optional encryption using a 40-bit shared-key RC4 PRNG algorithm from RSA Data Security. All data sent and received while the end station and access point are associated can be encrypted using this key. In addition, when encryption is in use, the access point will issue an encrypted challenge packet to any client attempting to associate with it. The client must use its key to encrypt the correct response in order to authenticate itself and gain network access.

Unfortunately, beginning with an internal study in 2000 (Walker, 2000) to a highly publicized study in 2001 (Borisov, Goldberg, & Wagner 2001), WEP has been shown to fall well short of accomplishing its security goals. Some of the problems of WEP that have been identified by researchers include the following:

WEP uses RC4, a synchronous stream cipher, but it is difficult to ensure synchronization during a complete session over the unreliable wireless link, leading to the use of a separate key for each packet—a clear violation of one of the most important requirements of RC4.

A very limited key-space is used, which is problematic since a separate key is needed for each packet.

802.11 does not provide any mechanism for sharing keys over an insecure channel.

There is no mechanism for a mobile to authenticate the network.

Checksum (CRC-32) used for integrity check is linear; thus, it is relatively easy to make undetected changes in the message.

Such weaknesses combine to result in a network that is vulnerable to several types of attacks and intrusions.

There are several ongoing efforts to secure the 802.11 network, one of which is the robust security network (RSN). In RSN, a recently proposed 802.1x standard (Institute of Electrical and Electronics Engineers, 2001) forms the basis for access control, authentication, and key management. In addition, a number of protocols such as extensible authentication protocol–transport layer security (EAP-TLS) (Aboba & Simon 1999; Diersk & Allen, 1999; Zorn, 1999) are being considered to provide strong authentication between the MS and the AP.

Timing and Power Management

Synchronization of all clocks within a BSS is maintained by periodic transmission of beacons containing timestamp information. In the infrastructure mode, the AP serves as the timing master and generates all timing beacons. Synchronization is maintained to within 4 ms plus propagation delay.

Timing beacons also play an important role in power management. There are two power saving modes defined: awake and doze. In the awake mode, stations are fully powered and can receive packets at any time. In the doze mode, it is unable to transmit or receive data and consumes very little power. A station must inform the AP that it is entering the doze mode. The AP does not send packets to stations in the doze mode, but instead buffers them for transmission at a designated time.

Comparison of 802.11b and 802.11a

Two advanced WLAN standards, 802.11b and 802.11a, were developed by the IEEE's 802.11 working group. At

Table 3 Features of 802.11b

Advantages	Disadvantages
2.4-GHz band is almost universally available.	Prone to interference from other devices that operate in the 2.4-Ghz band.

the MAC layer, they both use the CSMA/CA protocol. At the physical layer, 802.11b uses the DSSS radio transmission method and operates in the 2.4-GHz ISM band (see Table 3), while 802.11a uses orthogonal frequency division multiplexing (OFDM)—a more recent modulation scheme that is claimed to offer better performance at higher data rates (Bingham, 1990)—and operates in the 5-GHz UNII (unlicensed national information infrastructure) band (see Table 4).

802.11b—Offers data rates of up to 11 Mbps (see Table 5).

802.11a—Offers data rates of up to 54 Mbps due to higher carrier frequency and a more sophisticated encoding technology (see Table 5).

802.11g. While the 802.11g standard is yet to be finalized (draft standard was ratified in November 2001), it seeks to offer current users of 802.11b data rates up to 54 Mbps in the same 2.4-GHz band. Because 802.11b and 802.11g use the same frequency band, an 802.11b radio interface will work with an 802.11g access point albeit at an 802.11b rate. However, the physical range of 802.11g will be less than 802.11b, so higher concentration of access points will be necessary to obtain the full 54-Mbps rate throughout the service area.

Two mandatory and two optional modes are part of the draft standard. Use of OFDM (similar to one in use for 802.11a) is mandatory for data rates greater than 20 Mbps and support for complementary code keying (CCK) is necessary to ensure backward compatibility with existing 802.11b radios. The RTS/CTS mechanism described previously is used to ensure that both OFDM and CCK can coexist in the same 2.4-GHz band. An additional benefit of 802.11g is that since OFDM is already required for 802.11a, it is possible to build dual-band

Table 4 Features of 802.11a

Advantages	Disadvantages
Better performance in office environment (multipath reflection recovery). Higher data rates and less congestion in the UNII band. Some chip makers promise proprietary modes that will deliver up to 72 Mbps.	Limited number of channels available outside the United States.

Table 5 Comparison of 802.11b and 802.11a Technologies

Standard	802.11a	802.11b
Speed, up to	54 Mbps	11 Mbps
Range	300 feet	300 feet
Radio technology	OFDM	DSSS
MAC protocol	CSMA/CA	CSMA/CA
Frequency	5 GHz	2.4 GHz
Power management		Embedded power saving protocol

(2.4 and 5 GHz) radios without extra hardware complexity.

The popularity of IEEE 802.11a/b/g makes it a very serious challenger for 3G. It is being deployed in many public places to provide high-speed access to mobile users. The ease of deployment and maintenance of 802.11 is at the heart of its success. It is cheap (compared to 3G), robust, and simple to use, three cornerstones of any networking technology.

Bluetooth and PANs

The concept of personal area networks (PANs) has been introduced to enable wireless communication between devices in direct vicinity of a user. Examples of such devices are laptop computers, personal digital assistants (PDAs), cellular phones, printers, and photo cameras. PAN technology is characterized by low cost, low power consumption, and ad hoc network organization. Wireless PANs are being standardized in the IEEE 802.15 working group. The Bluetooth technology is being promoted as an industry standard for PANs, and is forming the basis for IEEE 802.15 standardization.

Bluetooth is a low-power, low-cost wireless technology for distances up to 10 m, at data rates up to 1 Mbps (and higher in newer versions of the standard) (Haartsen, 1998). It operates in the same 2.4-GHz ISM band as the IEEE 802.11b standard, using fast (1600 hops/s) FHSS. Bluetooth nodes are organized in so-called piconets, consisting of a master node and up to 7 slave nodes (that number is now being increased to 25 in the newer version of the standard). A slave can simultaneously be a slave or a master in another piconet, with a different frequency hopping sequence. This allows the construction of larger networks, called scatternets, where communication between nodes is carried out using multiple hops across piconets. Within piconets, the Bluetooth MAC uses polling to regulate access to the radio interface. A slave is only allowed to transmit a slot of data when polled by the master. Transmission of data from master to slave is considered as an implicit poll. Using this polling scheme, Bluetooth can provide both synchronous connection-oriented (SCO) links, e.g., for support of voice telephony applications, and asynchronous connectionless (ACL) links, e.g., for IP packet transfer.

Bluetooth products are rapidly being introduced into the market. Many new cellular telephone models are

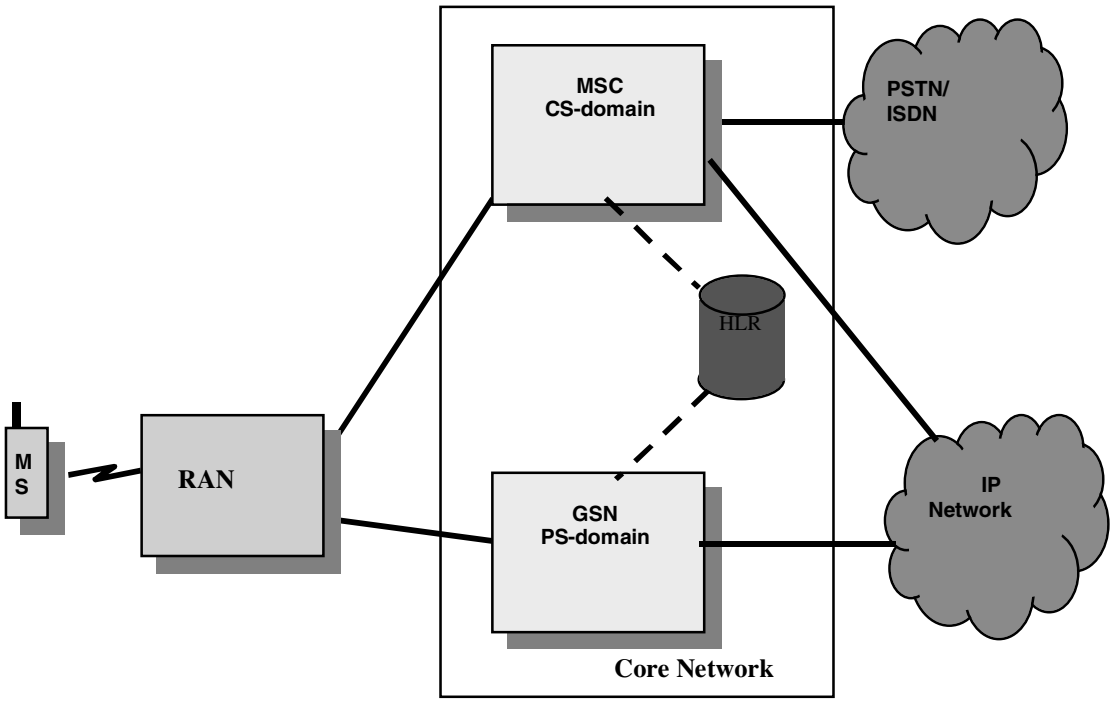


Figure 3: A 2.5G cellular network.

equipped with the technology as are some new digital cameras. Furthermore, products such as headsets, PCMCIA cards for laptops and PDAs are appearing on the market.

3G CELLULAR SYSTEMS

Second generation cellular networks are possessed with very limited data capabilities. Apart from special vertical services such as the SMS, these capabilities are restricted to circuit-switched data up to a data rate of 14.4 Kbps. 2G cellular networks have been extended later with enhanced data capabilities. The GSM standard, for instance, has been extended with a packet-switched mode, GPRS, where timeslots can be assigned toMSs on an on-demand basis. Also, an MS can combine several timeslots per frame, in order to achieve higher data rates (theoretically up to 171.2 Kbps). This GPRS is much better suited for the support of end-to-end IP-based services. For the migration to GPRS, only few changes are required in the GSM base stations. The core network consisting of circuit-switched MSCs, on the other hand, needs to be enhanced by a packet-switched network. In the core network the switches (MSCs) are augmented by specialized routers, the so-called GSNs. From this so-called 2.5 generation on, the core network of a cellular system consists of a circuit-switched part with MSCs, and a packet-switched part with GSNs (see Figure 3). The major change going from 2.5G to 3G cellular networks is a complete new RAN.

For 2G and 2.5G cellular systems, a number of incompatible standards and systems are used throughout the world. For 3G, the International Telecommunications Union (ITU) has set up IMT 2000 (International Mobile Telecommunications 2000) as a framework for worldwide wireless access by linking the diverse system of terrestrial and/or satellite based networks (ITU-T Recommendation Q.1701, 1999). The vision for IMT 2000 is to support advanced applications by providing higher data rates, from 384 Kbps global coverage to 2 Mbps indoor or low-range outdoor coverage. Further, the systems should be highly flexible, providing support for both applications that traditionally use circuit-switched networks and applications that traditionally use packet-switched networks. A wide range of date rates is to be supported, with a high granularity.

IMT 2000 comprises a number of cellular systems (De Vriendt, Lainé, Lerouge, & Xu, 2002). First, the CDMA2000 system is an evolution from the American CDMA system, IS-95. Further, a number of IMT 2000 systems are defined by (re)using the core network from GSM. Three different radio technologies have been defined for use with this core network. EDGE (enhanced data rates for GSM evolution) is an evolution from the GSM (and GPRS) technology, using new modulation techniques, to provide data rates up to 384 Kbps. The two other radio technologies have been developed in the context of UMTS (universal mobile telecommunication system), and are based on CDMA principles. The main difference between the two is the duplexing technique used. Duplexing the two directions of communication can be done in either frequency or time, resulting in an FDD and a TDD UMTS variant, respectively. The first one, which is currently most widely implemented, uses a technique called wideband CDMA (WCDMA), using 5-MHz carriers to provide data rates up to 2 Mbps. WCDMA and CDMA2000 both rely on CDMA technology, but differ from each other in various design and implementation aspects, such as clock and

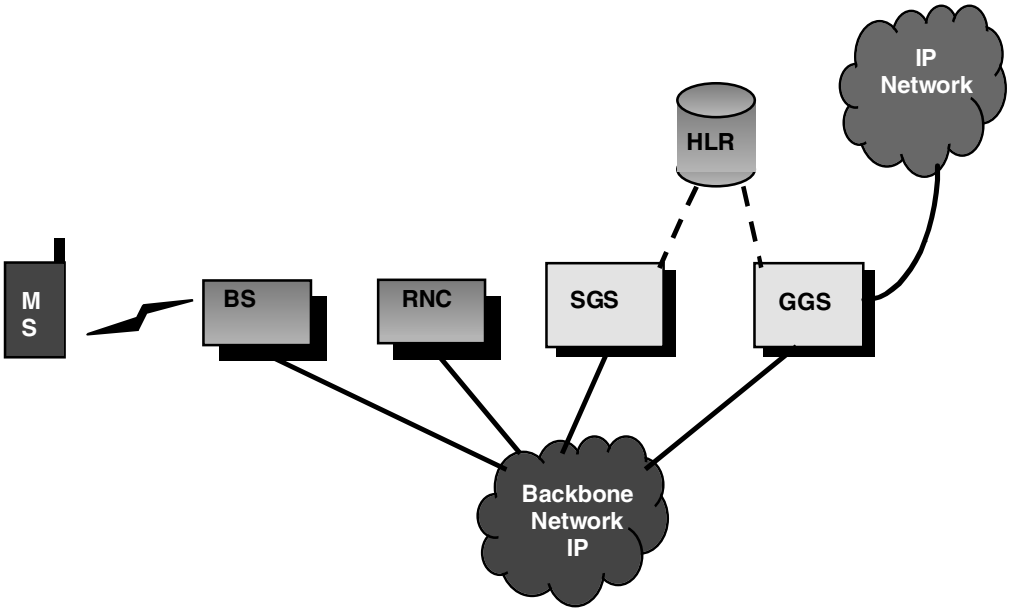


Figure 4: Architecture of a UMTS network.

chip rates, synchronization approach, pilot channels, and frame duration, and are incompatible with each other.

Both WCDMA and CDMA2000 provide the user with channels, either shared with other users or dedicated, with different maximum data rates, depending on the spreading factor used. It provides a wide range of channel bit rates, and is very well suited for bursty packet-based traffic. These radio standards offer quite some room for further improvement. New modulation techniques, more advanced scheduling, dynamic link adaptation, and multiple-input-multiple-output antenna techniques are some of the improvements currently under study (Honkasalo, Pehkonen, Niemi, & Leino, 2002). The high-speed downlink packet access channel (HSDPA) in WCDMA, with data rates up to 10 Mbps, has been standardized as a first step in this direction.

2.5G systems are widely available now throughout the world, although the use of these systems is not (yet) overwhelming. 3G systems are currently starting service in limited areas, especially in Japan (WCDMA), and are being expanded in functionality, performance, and coverage. Availability of terminals is still very limited.

CONVERGENCE OF IP AND CELLULAR SYSTEMS—TOWARD THE MOBILE INTERNET

It is expected that the future will show a convergence between the Internet and cellular systems. They will converge both in the services provided and in the technology used. The converged system will combine the wide range of horizontal services offered by the Internet with mobility and highly integrated services and devices offered by the cellular networks. As for the technology, it is expected that future systems will combine radio technology from cellular systems with switching and routing technology based on the IP protocol suite. Looking at a possible network architecture of a UMTS network (Figure 4), we see two distinct IP networks. One is the external IP network, the other an IP-based transport network. The external IP network is a network providing services and connectivity from remote end-users to the user of the MS. For the MS, the GGSN is a gateway to the external IP network, and provides the MS with an IP address valid on that network. IP headers of packets on their way from the MS to the external IP network (or vice versa) are only processed by the GGSN, not by all the intermediate nodes. From the point of view of the external IP network, the UMTS network is a single subnetwork.

The IP-based transport network is typically owned by the cellular operator, and provides connectivity between the various nodes in the UMTS network. These are, for the packet-switched core network, the SGSN and the GGSN, and for the RAN, the BSs and Radio Network Controllers (RNCs). The BS provides the main radio (physical layer) functions, whereas the RNC provides the higher layer functions, including radio resource management, and soft handover.

For the packet-switched core network, it has been specified that the SGSNs and GGSNs should be interconnected by an IP-based transport network. User-level IP packets are encapsulated and tunneled using the so-called GPRS tunneling protocol (GTP) on top of the transport level IP. GTP provides for the routing of the user level IP packets to the appropriate SGSN, using the database of the HLR, for mobility management functions. As such, GTP provides an alternative to MIP.

In the first release of UMTS, ATM is used as transport technology for the RAN. However, it is foreseen that IP will also be introduced in this part of the cellular network in the near future. The use of IP as transport technology in a RAN, besides enabling the operators to provide new packet-based services, is especially beneficial since it provides the means for statistical aggregation of traffic, which leads to increased transmission efficiency and

consequently reduced leasing costs. A radio frame is a short data segment coded/decoded and transmitted/received by the base station. These radio frames must be delivered from RNC to BS and vice versa in a timely fashion with limited delay. Otherwise, the BS or RNC will discard them. Due to the time constraints on the delivery of radio frames the majority of the traffic in an IP-based RAN can be considered to be real-time traffic. Seen from the architecture of a RAN and the nature of the transported data, the IP-based RAN has different characteristics when compared to traditional IP-networks. Typically, the wireline transmission in a radio-break access network contains a relatively high volume of leased lines. The fact that thousands of radio BSs are spread over a wide geographical area and are in general situated at large distances from the backbone typically results in high cost for the transmission links. Further, the majority of the traffic transported on the wireline transmission links used by the RAN is radio frames. This means that the traffic is very sensitive to delays and delay variation (jitter). Deploying resource management schemes in this environment is therefore essential.

The introduction of IP-based transport in the RAN indicates that an IP QoS-capable domain will have to be managed in the radio access network. Currently, DiffServ (Blake et al., 1998) as a scalable IP QoS architecture is the favorite one to be used in an IP-based RAN. The scalability is achieved by offering services on an aggregate basis rather than per flow and by forcing as much of the per-flow state as possible to the edges of the network, that is, to the edge nodes. In order to allow for dynamic resource management in DiffServ, an extension, called RMD (resource management in DiffServ), has been proposed (Heijenk, Karagiannis, Rexhepi, & Westberg, 2001). RMD extends the DiffServ architecture with new reservation concepts and features, such that the IP-based RAN resource management requirements are met.

These trends will lead to a cellular network architecture, where all nodes are interconnected using an QoS-enabled IP-based transport network. On top of this IP protocol, protocols for cellular specific functions, related to mobility and radio, will be running. Further, an end-to-end IP protocol will run on top of this to enable end-users to use IP-based services, and to connect to other end-users.

Currently, IP-based services over cellular networks are mainly best-effort type of services, both interactive, such as Web browsing, and background, such as e-mail downloading. For the near future, IP-based services might be extended with more streaming type of services for audio and video. A situation where all services, including conversational services, in a cellular network are IP-based is somewhat further away. Prerequisites for such a situation are very efficient header compression techniques, and a migration to a cellular network architecture where also signaling is IP-based, e.g., based on SIP. Efforts in these directions are being made in both research and standardization.

Cellular networks provide wide-area coverage for mobile users at moderate data rates. Besides cellular networks, other wireless systems are gaining popularity, in particular WLAN and short-range technologies. WLAN systems provide wireless ethernet extension to notebooks. WLANs based on the IEEE 802.11b standard (Institute of Electrical and Electronics Engineers, 1999) and operating in the 2.4-GHz ISM band are widely available in the market. They offer data rates up to 11 Mbps and have a range of 50 to 300 m. New products operating in the 2.4-GHz ISM and 5-GHz band offering data rates up to 54 Mbps are starting to appear on the market. These systems have been primarily designed for nomadic applications and, consequently, their support for mobility is very poor.

Short-range technologies, e.g., Bluetooth, have a more limited coverage, support lower data rates, and consume less power than WLANs (Haartsen, 1998). Operating in the ISM band, these technologies originally designed to replace the cabling connecting peripherals and other devices are also suitable for inexpensive communication between portable devices.

The above-mentioned technologies have been optimized for different applications. Even as they evolve, it is not expected that they will be replaced by a common multipurpose technology in the future. On the contrary, it is expected that the next generation wireless systems will integrate different and complementary technologies. Wireless devices able to operate with different radio interfaces will access the communication facilities using the "always best connected" paradigm. This means that a wireless device that has the choice will use that radio interface that it deems most appropriate for its purposes, e.g., highest performance or lowest cost.

In the future, a person might use multiple personal devices, such as laptop, phone, and organizer, that are mutually interconnected, forming a PAN (see Figure 5) using, for example, Bluetooth technology. These devices will all have one or more wireless interfaces. At a certain moment, the person might use his laptop and be connected to his company network via a wireless LAN interface. When moving out of the office, he may want to stay connected using a Bluetooth link between his laptop and phone, where his phone will act as an intermediate hop, via a UMTS link to the fixed network. In this scenario, the PAN acts as a moving network with multiple interfaces, which moves along different wireless APs to the fixed network, and which may merge with another moving network, e.g., the network in a vehicle.

The challenge in these scenarios is to achieve *seamless integration*, meaning that from the point of view of the application, switching from one network technology to the other is imperceptible and the level of security is maintained. This requires measures to be taken at different levels of the protocol stacks. Suitable techniques for supporting mobility and smooth transitions between different technologies and systems as well as in between private and public networks are under study (Wireless World Research Forum, 2001).

Even as wide deployment of 3G has been experiencing delays, 4G wireless technology is in an active research stage. 4G is intended to provide much of what 3G had originally envisioned, i.e., a broadband cellular service providing high-speed capacity at low cost, along with IP-based applications and services. Data rates of up to 20 Mbps are targeted, even as the MS moves at up to 200 km/h, and

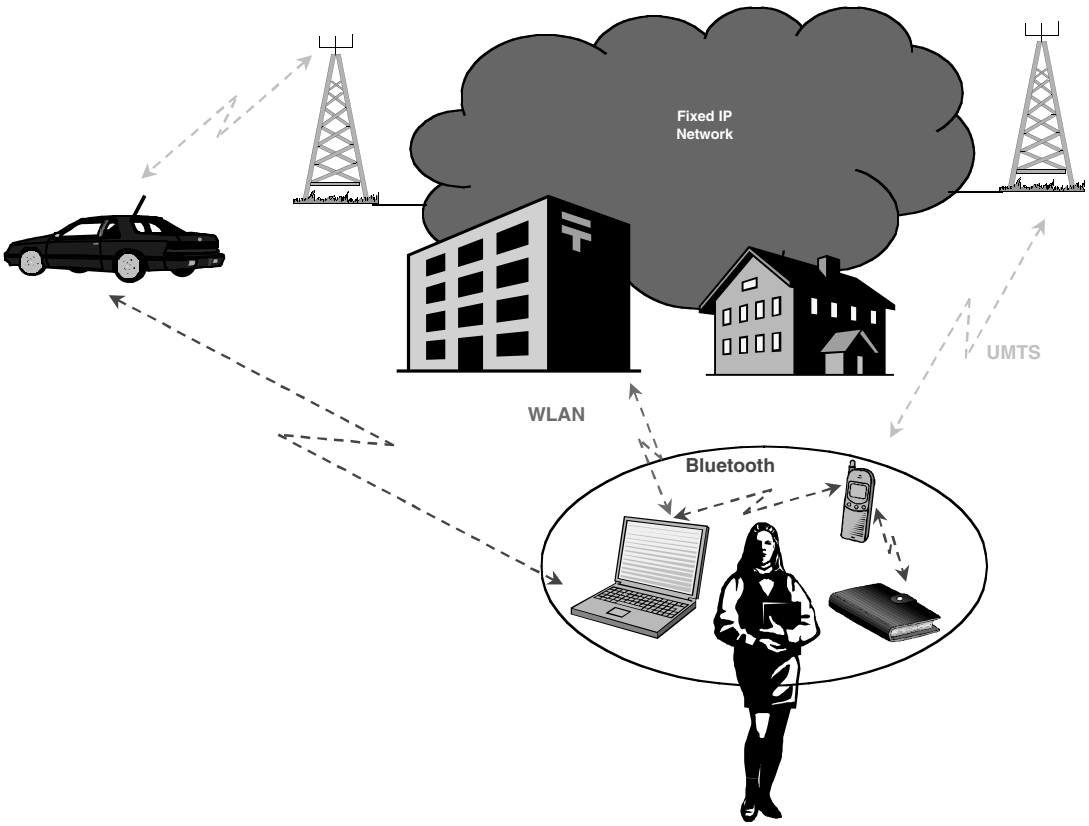


Figure 5: Depiction of a personal area network (PAN).

the entire network will use packet-switching techniques. However, because there is not a *single* 3G standard upon which to build 4G, there are significant challenges ahead.

To support such high data rates, significant advances in baseband processing are necessary. One of the most widely considered transmission schemes is multicarrier modulation (MCM), which uses many parallel subchannels to transmit information. MCM's advantages include better protection against intersymbol interference (ISI) and implementation efficiencies possibly using fast Fourier transform (FFT) techniques. At least two different types of MCM are being considered for 4G, including multicarrier CDMA (MC-CDMA) and OFDM (recall that OFDM is already used in several wireless LAN standards such as 802.11a and 802.11g). Regardless of the multiplexing scheme, techniques such as smart antenna and multiuser detection will become integral parts of the standard, as they enable the technology necessary to support the requirements of 4G. Additional information about 4G issues may be found in Lu (2002), Lu and Berezdivin (2002), and Mahonen and Polyzos (2001).

ACKNOWLEDGMENTS

We thank the following colleagues for their assistance with the authoring of this chapter: Jasmine Zan, Vlora Rexhepi, Georgios Karagiannis, and Sonia Heemstra de Groot. We also are very appreciative of all the insightful comments from the reviewers of this chapter and the guidance provided to us by Dr. H. Bidgoli.

GLOSSARY

ATM A network technology designed to transport cells, where unlike IP, cell sizes are fixed and circuit-switching circuits are used.

Autonomous system (AS) A network or a group of networks operated by a common administrator.

CDMA A digital cellular technology that uses spread-spectrum techniques to multiplex numerous concurrent connections.

Cellular network A wireless network that divides geographical regions into cells, to which subsets of available frequency bands are allocated.

Duplexing Transmission of data in two directions, from node A to node B and vice versa.

Encapsulation The process of embedding an additional network protocol on top of the existing one(s) possibly for the purposes of tunneling.

Hop A connection between two adjacent network devices such as a router.

Hypertext markup language (HTML) A formatting language used in creating documents for the World Wide Web.

Hypertext transport protocol (HTTP) A transport protocol used in the World Wide Web to deliver documents to browsers.

Internet A global network of computers and smaller networks.

Internet protocol (IP) A specification for data format and addressing scheme used on the Internet.

Intersymbol interference (ISI) Distortion of the received signal due to temporal spreading and interference of the transmitted pulses.

Medium access control (MAC) A protocol for ensuring that the common wireline/wireless medium is shared nicely among many potential users.

Node A device on the network, such as a workstation or a router.

Roaming The ability to establish and maintain wireless connection outside the home area of the user.

Router A device that performs routing.

Routing The process of moving packets from the source to the destination in a network.

Routing table A list in the router that matches the destination IP addresses to the physical ports that lead to the optimal path to the destination.

Short messaging service (SMS) A service available on many cellular networks that enables near real-time transmission of short text message among the subscribers.

SIP A signaling protocol used to set up and control interactive applications over the Internet.

Tunneling The process of transporting encapsulated data packets through foreign networks that may be using different network protocols.

Voice over IP (VoIP) A service that delivers voice communication over the Internet using IP.

World Wide Web (WWW) A large group of servers on the Internet that uses HTTP to transport files and documents, often formatted using HTML.

CROSS REFERENCES

See *Mobile Commerce; TCP/IP Suite; Voice over Internet Protocol (IP); Wireless Application Protocol (WAP).*

REFERENCES

3GPP TS 23.040. (2002, April). *Technical realization of the short message service (SMS), version 4.6.0.*

3GPP TS 23.140. (2002, April). *Multimedia messaging service (MMS); functional description; stage 2, version 4.5.0.*

Aboba, B., & Simon, D. (1999, October). *PPP EAP TLS authentication protocol* (RFC 2716).

Bakre, A., & Badrinath, B. (1995, May). I-TCP: Indirect TCP for mobile hosts. In *Proceedings of the 15th IEEE International Conference on Distributed Computing Systems* (pp. 136–143). New York: IEEE.,

Balakrishnan, H., Seshan, S., & Katz, R.H. (1995, December). Improving reliable transport and handoff performance in cellular wireless networks. *ACM Wireless Networks, 1*(4), 469–481.

Bender, P., Black, P., Grob, M., Padovani, R., Sindhushayana, N., & Viterbi, A. (2000, July). CDMA/HDR: A bandwidth efficient high speed wireless data service for nomadic users. *IEEE Communications Magazine, 38*(7), 70–77.

Bingham, J. A. C. (1990, May). Multicarrier modulation for data transmission: An idea whose time has come. *IEEE Communication Magazine, 28*(5), 5–14.

Black, U. (1997). *ISDN and SS7: Architecture for digital signaling networks.* Englewood Cliffs, NJ: Prentice Hall.

Blake, S., Black, D., Carlson, M., Davies, E., Wang, Z., & Weiss, W. (1998, December). *An architecture for differentiated service* (RFC 2475).

Borisov, N., Goldberg, I., & Wagner., D. (2001, July). Intercepting mobile communications: The insecurity of 802.11. In *Proceedings of the 7th International Conference on Mobile Computing and Networking* (pp. 180–189). New York: ACM SIGMOBILE.

Braden, R., Clark, D., & Shenker, S. (1994, July). *Integrated services in the Internet architecture: An overview* (RFC 1633).

Brady, P. T. (1971, January). Effects of transmission delay on conversational behavior on echo-free telephone circuits. *Bell Technical Journal, 50*(1), 115–134.

Brewer, E. A., Katz, R. H., Chawathe, Y., Gribble, S. D., Hodes, T., Nguyen, G., et al. (1998, October). A network architecture for heterogeneous mobile computing. *IEEE Personal Communications Magazine, 5*(5), 8–24.

Brown, K., & Singh, S. (1997, October). M-TCP: TCP for mobile cellular networks. *ACM Computer Communication Review, 27*(5), 19–43.

Caceres, R., & Iftode, L. (1995, June). Improving the performance of reliable transport protocols in mobile computing environments. *IEEE Journal on Selected Areas in Communications, 13*(5), 850–857.

Cai, J., & Goodman, D. J. (1997, October). General packet radio service in GSM. *IEEE Communications Magazine, 35*(10), 122–131.

Deering, S., & Hinden, R. (1998, December). *Internet protocol, version 6 (IPv6)* (RFC 2460).

De Vriendt, J., Lainé, P., Lerouge, C., & Xu, X. (2002, April). Mobile network evolution: A revolution on the move. *IEEE Communications Magazine, 40*(4), 104–111.

Diersk, T., & Allen, C. (1999, January). *The TLS protocol* (RFC 2246).

Doyle, J. (1998). *Routing TCP/IP* (Vol. I). Sebastopol, CA: Cisco Press.

Fielding, R., Gettys, J., Mogul, J., Frystyk, H., & Berners-Lee, T. (1997, January). Hypertext transfer protocol—HTTP/1.1 (RFC 2068).

GSM Association. Retrieved April 8, 2003, from http://www.gsworld.com

Haartsen, J. (1998). Bluetooth—The universal radio interface for ad hoc, wireless connectivity. *Ericsson Review, 75*(3), 110–117.

Hedrick, C. (1988, June). *Routing information protocol* (RFC 1058).

Heijenk, G., Karagiannis, G., Rexhepi, V., & Westberg, L. (2001, September). *DiffServ resource management in IP-based radio access networks.* Paper presented at the 4th International Symposium on Wireless Personal Multimedia Communications, Aalborg, Denmark.

Hinden, R. (1993, September). *Applicability statement for the implementation of classless inter-domain routing (CIDR)* (RFC 1517).

Honkasalo, H., Pehkonen, K., Niemi, M. T., & Leino, A. T. (2002, April). WCDMA and WLAN for 3G and beyond. *IEEE Wireless, 9*(2), 14–18.

How to make money on mobile services (2002). Strand Consult Publications. Retrieved May 13, 2003, from http://www.strandreports.com/sw494.asp

Huitema, C. (2000). *Routing in the Internet* (2nd ed.). Englewood Cliffs, NJ: Prentice Hall.

Instant messaging more popular than ever at work (n.d.). Retrieved May 13, 2003, from http://www.instantmessangers.com/site/news/im_more_popular_than_ever.htm

Institute of Electrical and Electronics Engineers. (1999). *Wireless LAN medium access control (MAC) and physical layer (PHY) specifications, IEEE 802.11.*

Institute of Electrical and Electronics Engineers. (2001). *IEEE 802.1X—Port based network access control.*

ITU-T Recommendation Q.1701. (1999). *Framework for IMT-2000 networks.*

Jakes, W. C. (1993). *Microwave mobile communications.* New York: IEEE Press.

Lu, W. W. (Ed.). (2002, March). Fourth-generation mobile initiatives and technologies. *IEEE Communications Magazine, 40*(3), 104–145.

Lu, W. W., & Berezdivin, R. (Eds.). (2002, April). Technologies on fourth generation mobile communications [Special Issue]. *IEEE Wireless Communications, 9*(2).

Mahonen, P., & Polyzos, G. C. (2001, December). European R&D on fourth-generation mobile and wireless IP networks. *IEEE Personal Communications, 8*(6), 6–64.

Mathis, M., Mahdavi, J., Floyd, S., & Romanow, A. (1996, October). *TCP selective acknowledgement options* (RFC 2018).

Mills, D. L. (1984, April). *Exterior gateway protocol formal specification* (RFC 904).

Mobile Application Part (MAP) Specification. (1997). ETS 300 599 (GSM 09.02), ETSI recommendation.

Mouly, M., & Pautet, M. B. (1992). *The GSM system for mobile communications,* Palaiseau, France: Cell & Sys.

Moy, J. (1998a). *OSPF anatomy of an Internet routing protocol.* Reading, MA: Addison-Wesley.

Moy, J. (1998b, April). *OSPF Version 2* (RFC 2328).

Nielsen/NetRatings First Quarter 2002 Global Internet Trends report (2002). Retrieved May 9, 2002, from http://www.nielsen-netratings.com/pr/pr_020509_eratings.pdf

Perkins, C. (1996, October). *IP mobility support* (RFP 2002).

Perkins, C. (1997). *Mobile IP: Design principles and practice.* Reading, MA: Addison Wesley.

Postel, J. (1980, August). *User datagram protocol* (RFC 768).

Postel, J. (1981a, September). *Internet protocol* (RFC 791).

Postel, J. (1981b, September). *Transmission control protocol* (RFC 793).

Postel, J., & Reynolds, J. (1985, October). *File transfer protocol (FTP)* (RFC 765).

Raggett, D., Le Hors, A., & Jacobs, I. (1999). HTML 4.01 specification, W3C recommendation.

Rekhter, Y., & Li, T. (1993, September). *An architecture for IP address allocation with CIDR* (RFC 1518).

Rekhter, Y., & Li, T. (1994, July). *A border gateway protocol 4 (BGP-4)* (RFC 1654).

Rosen, E., Viswanathan, A., & Callon, R. (2001, January). *Multiprotocol label switching architecture* (RFC 3031).

Schulzrinne, H., Casner, S., Frederick, R., & Jacobson, V. (1996, January). *RTP: A transport protocol for real-time applications* (RFC 1889).

Simpson, W. (1994, July). *The point-to-point protocol (PPP)* (RFC 1661).

Sklower, K., Lloyd, B., McGregor, G., Carr, D., & Coradetti, T. (1996, August). *The PPP multilink protocol (MP)* (RFC 1990).

Telecommunications Industries Association. (1991). Cellular radio-telecommunications intersystem operations, TIA/EIA/IS-41B.

Telecommunications Industries Association. (1993). Mobile-station base station compatibility standard for dual-mode wideband spread spectrum cellular system, IS-95.

Telecommunications Industries Association. (1996). 800 MHz TDMA cellular radio interface mobile station base station compatibility, IS-136A.

Telecommunications Industries Association. (2001). Physical layer standard for CDMA2000 spread spectrum systems, TIA/EIA/IS-2000.

Viterbi, A. J. (1995). *CDMA principles of spread spectrum communications.* Reading, MA: Addison-Wesley.

Walker, J. R. (2000, October 27). *Unsafe at any key size: An analysis of the WEP encapsulation* (802.11-00/362).

Wireless World Research Forum. (2001, December). *The book of visions 2001—Visions of the wireless world, Version 1.0.*

Yacoub, M. D. (1993). *Foundations of mobile radio engineering.* Boca Raton, FL: CRC Press.

Zhang, L., Deering, S., Estrin, D., Shenker, S., & Zappala, D. (1993, September). RSVP: A new resource reservation protocol. *IEEE Network, 7*(5), 8–18.

Zorn, G. (1999, January). *PPP LCP internationalization configuration option* (RFC 2484).

Wireless Marketing

Pamela M. H. Kwok, *Hong Kong Polytechnic University, China*

INTRODUCTION

Wireless technology began its growth in the 1980s and the rapid development of wireless products such as mobile phone, pagers, and other handheld devices has transformed the ways that people conduct business (Lamont, 2001). Surfing the Web or online services is another way to get product information and order merchandise from catalogs. . . . Firms can quickly update information at minimum "costs" (Boone & Kurtz, 2001, p. 25).

During the mid-1990s, the Internet became the main online direct-marketing channel that allows users to send e-mail and access public news, product information, and online shopping activities, among other things (Kotler & Armstrong, 2001). Nowadays, the development of high-speed wireless Internet, Web-enabled mobile phone handsets, and other wireless technologies has further sustained the application direct-marketing or one-to-one marketing in the wireless business world.

WIRELESS TECHNOLOGIES AND WIRELESS MARKETING

Advancement in wireless technology such as WAP, iMode, GSM, GPRS, UMTS, wireless LAN, WCDMA (see Table 1 for details) has enabled marketers to develop more effective and specific marketing (or one-to-one marketing) programs tailored to the needs of individual customers. To fully utilize these technologies, marketers must learn how to work with their customers to develop long-term mutual benefit relationships. Obviously, wireless marketing offers new business opportunities and threats to m-marketers. Marketers must understand the key dynamic developments in the wireless world to capitalize on business opportunities at the right time and take proactive strategies to eliminate potential risks.

What Is Wireless Marketing?

Wireless is defined as "radio communications" or a "radio receiver or transmitter" (*Oxford Advanced Learner's English–Chinese Dictionary*, 4th Edition, 1994). Microsoft Dictionary (1998, available on CD-ROM) defined wireless as "pertaining to, or characteristics that take place without the use of interconnecting wires or cables, such as by radio, microwave, or infrared." Marketing is regarded as "a social and managerial process whereby individuals and groups obtained what they need and want through creating and exchanging products and value with others (Kotler & Armstrong, 2001, p. 6). So wireless marketing

Table 1 Summary of Key Wireless Technologies

WAP	Wireless application protocol
i-Mode	A wireless service launched in Japan in spring 1999 by NTT DoCoMo. The service is accessed by a wireless packet network (PDC-P), and the contents are described in a subset of the hypertext markup language.
GSM	Global system for mobile communication
GPRS	General packet radio service
UMTS	Universal mobile telephone system (third-generation telecommunications system based on wideband code division multiple access)
Wireless LAN	Wireless local access network
WCDMA	Wideband code division multiple access

Source: Nokia (http://www.nokia.com).

refers to the exchange process (i.e., buying and selling) that is carried out by wireless means (e.g., mobile phone, m-commerce, wireless Internet, or any handheld wireless devices).

Major Services Are Provided by Third-Generation Technologies

The advancement in third-generation (3G) technology brings additional benefits to consumers. Multimedia functions of 3G consist of several media components which allow wireless delivery of video clips, still images, and music. It can be "interactive and distributional" (i.e., users may have subscribed to certain distributional applications, and receive only those subscribed services). The maximum speeds of GPRS can reach up to 171.2 Kbps faster data transmission speed or download time. So GPRS enables interactive visual display (i.e., MMS [multimedia message service]) and enhancing communications quality.

Besides these benefits, 3G technologies deliver the following value-added wireless services as well:

1. Internet access (e.g., users can download ring-tones, music, cartoon characters, real-time events)
2. Location-based application (e.g. users can send promotional messages to specific groups of customers)
3. Simple games (e.g., users can play card games and crosswords)

Factors Influencing Adoption of Wireless (or Mobile) Internet

The decision-making process is complex, and there is no perfect formula to explain customers' purchase behavior. An industry report published by the International Telecommunication Union (ITU, 2002a) in September 2002 reported that the adoption of mobile Internet service might be due to the following factors:

- Application of mobile multimedia services such as video clips or still images, video, and music through 3G technologies
- Availability of Internet-enabled handsets with affordable prices
- "Unrestricted and nonproprietary" mobile Internet content
- Simple billing systems or models for both voice and data transmission services

Mobile communication and Internet access have been the key drivers for consumer telecommunications services in recent years. In 2002, the ITU (2002b) indicated the top 20 mobile and Internet markets worldwide (Table 2).

MOBILE PENETRATION IN THE UNITED STATES AND EUROPE

e-Marketer reported the interactive survey conducted by Telephia and Harris (Figure 1). The survey indicates that Greenville had the highest mobile phone's penetration (i.e., 71%) among 35 major U.S. cities by December 2002. It is closely followed by St. Louis (69%). Forrester Research study in March 2002 reported that Finland, Norway, Sweden, and Italy have highest mobile penetration (Figure 2).

Marketing Opportunities and Wireless (or Mobile) Internet

An analysis of Internet usage and projections covering more than 50 countries, carried out by Computer Industry Almanac in March 2002, reported that there would be 1.12 billion Internet users worldwide by the end of 2005, and wireless Internet users will reach 48%. "The wireless Internet will take off when always-on service and useful content for the small displays of wireless devices are available," predicts Dr. Egil Juliussen, author of the report (Computer Industry Almanac, 2002). Access to the Internet through personal digital assistants (PDAs) with multiple functions (e.g., built-in Internet access, digital camera, music player, and scanner) is expected to increase in developed countries.

Interestingly, these wireless devices are expected to be considered the primary devices for Internet access in countries with low Internet penetration rates. The report suggests that 41.5% of the worldwide population will use the wireless Internet by 2004. In some developing markets such as the Philippines, the penetration rate of mobile phone services was three times higher than fixed-line penetration at the end of 2001.

Table 2 Top 20 Mobile and Internet Index Rankings, Worldwide

ECONOMY	MOBILE/INTERNET SCORE (per 100)	RANKING
Hong Kong, China	65.88	1
Denmark	65.61	2
Sweden	65.42	3
Switzerland	65.10	4
United States	65.04	5
Norway	64.67	6
Korea, Rep. of	63.42	7
United Kingdom	63.00	8
Netherlands	62.25	9
Iceland	62.03	10
Canada	61.97	11
Finland	61.22	12
Singapore	60.58	13
Luxembourg	58.58	14
Belgium	57.80	15
Austria	57.7	16
Germany	55.53	17
Australia	55.40	18
Portugal	55.13	19
Japan	54.94	20

Note: From the International Telecommunications Union Mobile/Internet Index included in the *Internet for a Mobile Generation Report.* The index measures how each economy is performing in terms of information and communication technologies (ICTs) and captures how poised the country is to take advantage of future ICT advancements. The index covers 26 variables sorted into three groups: infrastructure, usage, and market structure. These three components combine for a score between a low of 0 and a high of 100. The table is taken from the Statistical Annex to the Report, which provides comprehensive data on network and service development for more than 200 economies. ©International Telecommunications Union (ITU), 2002. Reprinted with permission.

Richter and Mar (2002, p. 128) reported a recent IDC forecast that the estimated average growth of global spending on information technology (IT) would be about 10 to 11% between 2002 and 2005. The forecast report showed that the Asia Pacific region has the highest growth

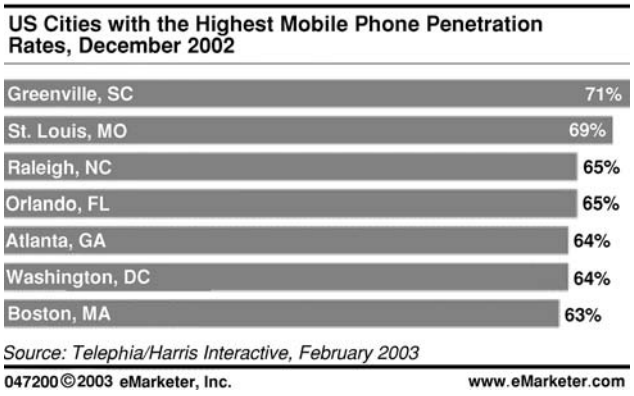


Figure 1: U.S. cities with the highest mobile phone penetration rates, December 2002. Source: Telephia/Harris Interactive, February 2003. ©2003 eMarketer, Inc.

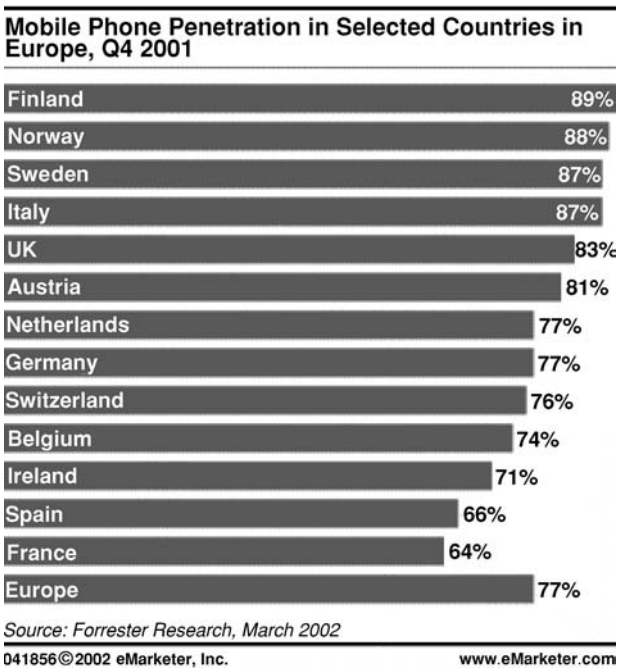


Figure 2: Mobile phone penetration in selected countries in Europe, fourth quarter 2001. Source: Forrester Research, March 2002. ©2002 eMarketer, Inc.

potential for IT spending (i.e., 11%) compared with other regions. Moreover, according to the forecast, extraordinary growth of about 30% is expected in countries such as India, China, Turkey, Egypt, and the Philippines.

For example, the penetration of value-added services of short message service (SMS) is popular and successful in the Philippines. Although that country has about 10 millions subscribers, 100 millions SMS messages are sent every day. In fact, making e-jokes through SMS has become part of Filipino culture. The subscribers enjoy circulating e-jokes among themselves. In addition, the second-generation handsets are widely available and supported by a competitive pricing strategy (about 2¢ U.S. per SMS message) and a simple billing system (Ritcher & Mar, 2002). SMS is simple to adopt and adaptable for low-bandwidth services (i.e., always online).

E-entertainment (i.e., e-music, e-video, interactive TV, e-gambling, etc.) has been regarded as the main driver behind wireless technologies. The demand for e-entertainment service has led to new business opportunities for hardware manufacturers (e.g., mobile phone and components suppliers), service operators, and software and network suppliers.

Another emerging market is China, where mobile phone service subscribers are expected to exceed fixed-line subscribers by March 2003. Furthermore, China's current 5-year plan indicates that telecom service revenues will increase at twice the rate of the gross domestic product.

This implies the growing importance of wireless Internet services, particularly in those countries with a low Internet penetration rate. On the other hand, marketers can make use of this opportunity to develop creative wireless

marketing strategies to foster customer relationships and cultivate new business opportunities—at the right time, in the right place.

Will Wireless (or Mobile) Substitute Fixed-Line Internet?

Mobile Internet will not be treated as a substitute for the fixed-line Internet (ITU, 2002a, p. 3), although 3G is considered a global wireless medium. This is because mobile Internet services can be supported by different wireless technologies such as wireless local area networks (LANs), short-range connectivity technologies, fixed broadband networks, and so on. From the user's perspective, the following issues should be taken into considerations:

- Display quality is limited by the screen size of mobile devices.
- Functions provided by a keyboard are superior to those of a mobile device keypad.
- Fixed-line broadband has brought speedy connection benefits at reasonable prices.

Nowadays, the majority of display screens provided by wireless devices (e.g., PDAs and Web-enabled mobile phones) are too small for viewing long documents such as a contract or a book chapter. In addition, more and more consumers in developed countries have adopted fixed-line broadband Internet access services at home. For example, going online through fixed-line broadband Internet services at home has become part of the daily lives of Hong Kong residents. Although mobile messages through SMS and multimedia messaging service (MMS) provide "instant" and "interactive" benefits, they may not be able to substitute for land-line access at this time; nonetheless, they offer other benefits.

M-COMMERCE AND MARKETING

What Industries Benefit Most From M-commerce?

A new market space has been created by Web-enabled mobile phone services. The interactive features, always-on connections, and tailor-made contents benefit following industries (Lamont, 2001):

- M-banking (e.g., online stock trading—making enquiries and money transfer transactions): M-banking offers convenient and "instant" benefits through both voice and data communications. It allows customers to use their mobile phone services for online stock trading, checking balances, making money transfer, and so on anytime and anywhere. Security issues have been a key obstacle for influencing consumers' adoption of mobile-banking services.
- M-entertainment services (e.g., real-time sports, games, live-entertainment events, entertainment on Internet and music): The unique benefit of "real-time" enables subscribers to enjoy different types of live-entertainment services anytime and anywhere.

M-commerce and Information Services

M-commerce is the word used to describe the consumption of Internet services or the purchase of goods via a wireless link (Korhonen, 2001, p. 433). M-commerce is expected to grow continuously in the coming years. Furthermore, "everything available via the fixed Internet will also be available via the wireless Internet" (Korhonen, 2001, pp. 434–435). Wireless technologies allow speedy and real-time data transmission benefits. As a result, information services in m-commerce may apply to the following:

1. Real-time broadcasting services (e.g., on television, radio)
2. Location-based advertisement (e.g., a list of restaurants offering special discounts in a particular location targeting specific customers); it is suggested that the list should cover only those restaurants located within half a kilometer of the location of the handset
3. Purchase of consumer goods through mobile phone handsets
4. Paying for parking through SMS notices (e.g., sending reminders)
5. Conducting transactions with the parking meter through Bluetooth technology

Database Marketing in M-commerce

Database marketing can be applied to m-commerce in the following areas (adapted from Kotler & Armstrong, 2001, p. 626):

1. Identify business prospects: Information obtained from enquiries (e.g., through counter services, the Internet, e-mail, or mobile phone) and business reply cards enable companies to identify potential customers as well as business opportunities
2. Deciding which customers should receive a particular offer: Companies can tailor an offer (e.g., sending an e-coupon to a customer through mobile phone) according to the customer's preferences and needs.
3. Deepening customer loyalty: Companies can make use of database information to identify customers' interest by discovering their preferences and purchasing patterns. Furthermore, marketers can please customers by sending small gifts on special days. For example, a company could send an SMS birthday greeting, a birthday ring-tone download, or a discount e-voucher through mobile phone services before or on the customer's birthday.
4. Reactivating customer purchases: A database enables marketers to stimulate sales from potential customers such as replacement of a mobile phone handset, upgrading of a service plan, or purchase of a package or of complementary mobile phone products or services.

What Must Marketers Do to Be a Successful M-marketer?

Lamont (2001) proposed several criteria (Table 3) for being a successful m-marketer. In sum, m-commerce

Table 3 Criteria for Being a Successful M-marketer

M-MARKETING STRATEGIES	DETAILS
Product	Introduce miniature information appliances with unique interactive content to m-commerce customers
Price	Offer both commodity and higher value for money as marketing managers divide m-commerce customers into those who do virtually everything online and those who prefer personal services from telecom, content, and financial service providers.
Promotion	Provide customers with value-added intangible product attributes that are included as part of their smart handheld devices. For example, e-discount couples and free ring tone downloads
Segmentation	Divide like groups of people across national frontiers into those who have the income, are the correct age, live in the right neighborhoods, and belong to modernizing ethnic groups as candidates for the purchase of miniature information appliances, third-generation telecom services, and interactive Internet content.
Targeting	Assemble smaller like groups of people who are bound together by profession, skills, or personal tastes, habits, or values.
Positioning	Match possible online Internet products with probable customers; the former offers the latter enhanced customer relationships to try out m-commerce and the mobile Internet.

From *Conquering the Wireless World—the Age of M-Commerce,* by Douglas Lamont, p. 279. Oxford, UK: Capstone, 2001.

operators must offer unique, personalized services to their customers that offer value for the money. More important, m-commerce requires marketers to adopt a proactive approach and be sensitive to their marketing environment to pursue new business opportunities in the changing wireless world. M-marketers have to recognize that the product life cycle of portable wireless devices (e.g., PDAs and mobile phone handsets) is much shorter than it is for other consumer durables in technology sector.

Merits of Reconfiguring the Value Chain

The term "wireless world" implies global business perspectives and death of physical distance, and it offers convenience benefits to the services subscribers. "Businesses now have to compete on the basis that the world is a single market. Under such circumstances, the two keywords that are expected to lead corporations to a stable growth are "efficiency" and "creativity" (Richter & Mar, 2002, pp.142–143). Consequently, marketers need to "create new value through the wireless Internet services (Lamont, 2001, p. 55). Lamont summarized three merits of reconfiguring the value chain:

1. Firms can recognize and identify decision opportunities across the industrial world for wireless telecom services.
2. Firms can create better alternatives for making good decisions either through alliances and partnerships or through direct investments (or both).
3. Firms can establish a set of competitive principles for the firm as they seek to conquer the wireless world in the age of m-commerce.

WIRELESS CONSUMER BEHAVIOR

Key Impact of Mobile Number Portability (MNP) or Wireless Number Portability (WNP) on Customer Loyalty

The policies of MNP or WNP allow subscribers to carry their existing mobile telephone numbers to other network service providers. It enables subscribers to switch service provider more easily. Mobile phone subscribers are price sensitive and tend to keep on seeking service packages offering better value for money. This has had a significant impact on customer loyalty and churn rate. Marketers have to put more effort into cultivating customer loyalty and building long-term customer relationship to prevent high churn rate. Obviously, it costs companies more to acquire new customers than to keep their existing ones.

Unfortunately, MNP may disrupt subscribers' privacy when telemarketing calls are received through mobile phones. As a result, m-marketers need to consider permission marketing seriously and develop a strategy for pursuing consistent branding image to establish a brand community for fostering customer loyalty and reinforcing customer relationship.

Brand Community and Customers' Loyalty

Brand community exists because "brands connect consumers to brands, and consumer to consumer" (Muniz & O'Guinn (2001, p. 418). These social groups have a high propensity to be reasonably stable and devoted to the brand. Furthermore, members of a brand community tend to be "committed," "conscientious," and "passionate" toward the brand (Gruen & Ferguson, 1994).

Consequently, it facilitates marketers to strengthen their branding strategies because the Web enable firms to communicate instantly with each customer.

Consumer Decision-Making Process for Wireless

Adoption of wireless services can be divided into five stages (Steuernagel, 1999):

1. Awareness stage—information obtained from television advertisement, mass media, observation of users.
2. Interest stage—attention and interest are stimulated by multiple exposures to users and ads.
3. Evaluation stage—perceived benefits (e.g., job or lifestyle, emergency, security), making enquiries for different models of handsets.
4. Trial stage—through friends or promotional demonstrations.
5. Adoption stage—customer adopts wireless service and continues tracking the cost—benefit relationship of the adoption.

Segmentation Variables for Wireless Consumer Markets

Three major segmentation variables have been identified for wireless consumer markets (Lamont, 2001):

1. Demographic—generation (e.g., Generation Y), income groups (e.g., high disposable income groups who can afford to pay for the wireless services), occupation (e.g., professional athletes).
2. Psychographic—lifestyle (e.g., high mobility, enjoying high-tech life, people who prefer to do everything online), personality (e.g., personal value).
3. Behavioral—benefits (e.g., seeking speed, convenient services, etc.), loyalty (through brand community).

Permission Marketing and Customer Relationship

Godin and Peppers (1999) indicated the merit of permission marketing is that it allows both parties to enjoy "mutually beneficial dialogue" without worrying about the privacy and legal issues. "Permission Marketing has been around forever (or at least as long as dating), but it takes advantage of new technology better than other forms of marketing. The Internet is the greatest direct mail medium of all time, and the low cost of frequent interaction makes it ideal for Permission Marketing" (Godin & Peppers, p. 51). Wells, Burneet, and Moriarty (2003, p. 23) documented that this concept because many advertising messages are regarded as interruptive. As a result, permission marketing should be taken seriously as a means to foster good customer relationships and "mutually beneficial dialogue." This concept includes the following three principles:

1. The consumers (or audience) are in charge of the process.
2. The consumers (or audience) have agreed to receive promotional messages.
3. The consumers willfully sign up (i.e., willingness to participate).

Because consents are obtained from customers before sending promotional messages, the perception of invasion of personal privacy can be reduced. More important, customers are given an opportunity to control the process of marketing communication activities such as accepting or rejecting an SMS promotional message. This is an effective and efficient way to attract customers' attention and acceptance to the advertising messages. It also enables marketers to foster good customer relationships without invasion of personal privacy or sending offensive SMSs.

Participation Marketing and Customer Relationship

This concept goes beyond permission marketing, according to Alan Rosenspan (2000), an expert in direct marketing. It covers the following five principles (as cited in Wells et al., 2003, p. 24):

1. You really understand your customers (i.e., needs and preferences), for example, by collecting feedback from customers through marketing research, internal sales team, enquires, and customers' complaints.
2. You provide feedback at every opportunity (i.e., to show your concern and customer care), for example, by making a call to seek subscribers' opinions of new value-added services.
3. You involve customer and prospects as much as possible (i.e., stimulate customers' participation to build loyal customers, e.g., inviting customers to participate into charity events or public relations activities.
4. You market on customers' schedules, for example, by making promotional offers, such as sending coupons or e-gifts (e.g., free download services) on customers' birthdays.
5. You make customers feel vested in your success.

Ethical and Privacy Issues in Wireless Marketing

Similar to wired marketing (i.e. marketing activities conducted through fixed telephone lines), wireless marketing can annoy and interrupt customer's privacy. Ethical and privacy issues (e.g., irritation, fraud, invasion of personal privacy) should be taken into consideration when developing the marketing strategies to foster long-term customer relationships. Some examples of ethical and privacy issues that can arise in the mobile environment are as follows:

1. Irritation—sending out too many, too long SMS promotional or advertising messages to potential customers for pursuing location marketing strategies.
2. Fraud—making false claims such as "today is the last day for the special offer" when it actually isn't, overcharging customers on their bills, giving "misleading"

or "unclear" product information to potential customers.
3. Invasion of personal privacy—conducting telemarketing activities too early in the morning or too late at night, disclosing personal data to a third party for commercial or fundraising purposes.

Consequently, the concept of permission marketing should be encouraged and cultivated throughout the organization. In fact, companies need to set good practices and an effective system to monitor these consumer issues.

Consumer Protection

Wireless technologies have gone through a significant development period; however, the legal protection for m-commerce is still in its infancy. From an e-commerce perspective, OECD's (Organization for Economic Cooperation and Development) Consumer Protection Guidelines for e-Consumers (December, 1999) has indicated clearly about the possible consumer legal protection for e-commerce. The guidelines attempt to draw government, business bodies and consumer groups' attention for developing national and global awareness of consumer protection laws.

An ITU report (2002a), published in September 2002, proposed the creation of special legal protection for mobile consumers for the following reasons (OECD's guidelines can be regarded as a foundation for the legal protection of m-commerce):

1. Mobile devices are classified as high-speed terminals.
2. Users of mobile devices are numerous and always have limited usage experience.
3. Users usually have limited technical and legal knowledge of m-commerce.
4. The display screen for display on mobile devices is limited (it is difficult to display contract details).
5. Mobile devices have limited keypad functions (e.g., absence of "cancel" or "undo" keys) for making corrections.

MARKETING MIX FOR WIRELESS

Product

Product Characteristics

Lamont (2001) suggested that marketers should make use of brand name to differentiate themselves from their competitors. He also has identified the following product or service characteristics for wireless industry:

- Effective mobile devices should have interactive and instant features with unique content through wireless Internet. This content should be created for a particular mobile device (e.g., PDA or mobile phone).
- SMS (the "killer application" for mobile phones), MMS services, m-sports, m-commerce, m-entertainment, and m-banking are expected to be the major value-added services for mobile devices.
- New products develop on mini handsets and handheld mobile devices such as Web-based phones, WCDMA, and GPRS ("always on connection").

New Product Attributes in M-commerce

Dynamic wireless technology has changed the ways that people conducting their business, it also influences consumers' daily lives and their consumption behavior. The following summary of new product attributes for the wireless products and services is adapted from Lamont (2001, p. 48–49). Wireless devices can or soon will be able to

- find a parking space for your car,
- search for the best bargains for your family (e.g., where to eat and shop),
- transfer medical records to your physician,
- permit mothers to breast-feed their babies on time and at work,
- scan inventory and close a sale,
- download music recommended by your friends,
- chat with your friends,
- read your e-mail,
- pay bills while you commute to the city,
- transfer funds from your checking to your savings account,
- arrange travel and other services,
- conduct banking services (e.g., enquiries, transfer payment),
- watch entertainment programs (e.g., sports),
- take photos, and
- listen to the radio.

Price

In the early days of widespread Internet use, an initial pricing strategy for the fixed-lined Internet was charged according to time spent online. Later, flat-rate schemes or plans for different user groups were offered. Pricing for mobile (or wireless) Internet use is expected to follow the pattern of the fixed-line Internet (ITU, 2002a). Today subscribers are charged according to their rate of consumption (i.e., per minute) to browse the wireless Internet. Unfortunately, the wireless industry is competitive, and marketers cannot control their prices. Customers play a more active role in influencing the pricing strategies of wireless industry because of a high level of transparency in the wireless business world.

Pricing 3G

There are different types of services provided by 3G mobile devices, such as data services and real-time Internet game applications, so the billing system is more complex than it was 2G devices. It is expected that different tariffs will be adopted because some wireless services such as e-mail can accept delay, but other services such as Internet game applications cannot. As a result, managers need to consider what happens if a real-time Internet game application suffers from lower service quality because of net congestion, for example. A billing system must be able to solve this kind of problem automatically (Korhonen, 2001). Above all, Lamont (2001) indicated that customers needed to be educated about the merits of the wireless Internet (e.g., anytime, anywhere, voice and data communications) to ensure that potential customers of 3G are

willing to pay for different service rates and a smart billing systems is developed to solve some basic technical problems of Internet surfing (e.g. net congestion).

Prepaid Versus Postpaid

A prepaid concept is defined as customers who do not have subscriptions with a mobile operator but buy airtime for their SIM (Subscriber Identity Module) cards in advance and then use the phone as long as there is credit in their accounts. When their airtime runs out, they purchase more (Korhornen, 2001). SIM card is a small chip card kept inside a mobile phone. It serves as a network access card and it is similar to a tiny computer. It provides storing capacity for phone numbers, text messages, and value-added services. The advantages of prepaid service for the provider are as follows:

- There is no need to carry out a credit check on new customers.
- Customers with imperfect credit information can adopt mobile phone service under the prepaid scheme.
- There is no need to bill prepaid customers.
- Customers may never uses the services they pay for.

The advantages of prepaid service for users are as follows:

- Users remain anonymous in prepaid (hardware and software operators).
- Users have more freedom to decide where to buy a handset and SIM.
- There is no contract arrangement, and customers have more freedom to switch network operators.
- There is no monthly subscription charge, so light users can be attracted (e.g., keeping a mobile device for emergency purpose only).
- It is an efficient way for customers to control costs (automatically limits their phone service usage).

Disadvantages of prepaid service for the provider are the following:

- From the operators' perspective, there are security problems, because it is difficult to identify fraudulent users.
- The anonymous nature of prepaid service does not allow marketers to employ data mining or database marketing techniques (cannot identify their needs and characteristics).

The disadvantages of prepaid service for the customer are the following:

- Prepaid calls are more expensive than postpaid SIMs.
- Prepaid service may have an expiration date (unused airtime cannot be carried forwarded).

In Finland, the network usage is relatively inexpensive, and thus the prepaid with no monthly fee would not offer any particular benefit. Consequently, the demand for the prepaid is limited in that country. In Italy and Portugal, however, some operators have more than 80 prepaid scheme customers.

Distribution or Sale Channels

As discussed earlier, the wireless Internet is not going to replace fixed-line Internet in the immediate future, so the wireless distribution channel is regarded as a supplementary to existing sales channels (e.g., it could be applied to internal direct sales force, dealers, retailers, and direct marketing, etc.). Direct marketing is covered in more detail in the Promotion section.

Internal Sales Force—Telemarketing

The internal sales team of service providers can pursue telemarketing to individual potential customers. Sufficient training and retraining programs must be provided to all sales people to carry out effective telemarketing activities, however. In addition, motivation programs (e.g., incentives, commission, or bonus schemes) for rewarding outstanding salespeople and employees should be considered and integrated into the company's policies.

Middleman (e.g., Dealer, Retailers)

Service providers use middleman to attain a wider coverage of the marketplace. The middlemen play an important role in facilitating business transactions for both business and consumer markets.

Good channel relationship management is a key issue for pursuing success in channel management. To eliminate channel conflicts, however, the m-channel (mobile channel) should only be considered a supplementary mode to the traditional channel of distribution. Generally speaking, the channel of distribution can be divided into intensive, exclusive, and selective distribution. Because mobile phone service is perceived as a commodity, intensive distribution (i.e., as many outlets as possible; Kotler & Armstrong, 2001) should be considered.

Promotion

There are several promotional mix tools used in marketing strategy. With respect to m-commerce, advertising, public relations, and direct marketing are recommended.

SMS Advertising

SMS refers to the short message services provided by mobile network providers. SMS allows m-marketers to send electronic coupons and customized promotional messages to individual customers. Advisor.com (2002, Document No. 09546) reported a recent global study by HPI Research Group (marketing specialists) on the behalf of Nokia showing that a majority (88%) of customers like this mobile advertising method (i.e., sending advertising messages or e-coupons via SMS). The study covered 11 countries including Brazil, Denmark, Germany, Italy, Japan, Korea, Singapore, Spain, Sweden, the United Kingdom, and the United States.

Public Relations

Service providers can use public relations to build brand or corporate image. Public relations is defined as "building good relations with the company's various publics

Table 4 Real-Time Impact of Wireless Internet on the 4 Ps of Marketing

PRODUCT	Real-time wireless Internet transactions force products to become commodities, core assets to become peripheral, and valuable assets to become loss leaders. Recognize that goods and services from wireless content providers tend to be the same in the minds of the buyers. Marketers must create competitive transparency to succeed.
PRICE	Real-time m-commerce transactions base prices on demand at the time of sale, and these prices change continuously. Deliver the lowest possible prices and minimum transactions costs for goods and services from content providers. Users demand this competitive pricing structure in return for maximum purchases. Marketers must establish financial nakedness to succeed.
PLACE	Real-time wireless Internet transactions give all sellers power in the channel of distribution. Provide the maximum number of suppliers and minimum inventory levels. Customers insist on success in fulfillment or the delivery of accurate orders. Marketers must put in place distribution exposure.
PROMOTION	Real-time m-commerce transactions push well-established brands out of the market and into the dead brands society. Increase marketing openness as the norm for customers who want promotion and advertising to provide information rather than new entry barriers. Marketers must convert traditional push advertising to pull promotions that are targeted to specific individuals through permission marketing.

From: *Conquering the Wireless World—the age of m-commerce,* by Douglas Lamont, 2001, p. 28. Oxford, UK: Capstone.

by obtaining favorable publicity, building up a good corporate image, and handling rumors, stories, and events" (Kotler & Armstrong, 2001, p. 512).

Public relations consists of different types of promotional activities such as press releases for launching new products or services and holding charity events, seminars, or conference activities. Customers tend to believe news from public relations' events or activities more than they believe advertising messages. As a result, an effective public relations program enables a company to build a good corporate image in a cost-effective manner. Obviously, it also allows marketers to strengthen customer relationships through interactions with existing and potential customers.

Direct Marketing

There are five major tools in direct marketing: (a) direct mail, (b) catalogs, (c) telemarketing, (d) direct-response advertising (Wells et al., 2003, p. 407), and now (e) the Internet. The wireless technologies help marketers to collect more detailed information about their customers' needs and buying behavior. Thus, it enables marketing to tailor-made direct marketing strategies based on customers' individual's needs. Direct marketing tools include the following:

1. Direct mail—SMS and MMS through mobile phones.
2. Catalogs—e-catalogs provided in Web pages, SMS, or m-coupons of promotional offers sent through mobile phones.
3. Telemarketing (message must be simple)—call customers directly through mobile phone for information promotional offers.
4. Direct-response advertising (to obtain action-oriented objectives)—interactive advertising messages hosted in m-commerce or the wireless Internet, which are sent directly to customers through mobile phone.

In addition, service providers may apply either push or pull promotion strategies for to pursue their promotional objectives (e.g., to create a level of awareness). Kotler and Armstrong (2001, p. 531) defined the two strategies as follows:

Push strategy is a promotion strategy that calls for using the sales forces and trade promotion to push the product through channels.

Pull strategy is a promotion strategy that calls for spending a lot on advertising and consumer promotion to build up consumer demand, which pulls the product through the channels.

Real-Time Impacts of the Wireless Internet on the Four Ps of Marketing

Lamont (2001) summarized the "real-time" impact of wireless Internet on the Four Ps of marketing (product, price, place, promotion) as detailed in Table 4.

Wireless Marketing and Examples of Business Models

Companies adopt different business models based on their strengths and their market situations. The followings are some examples of business models for wireless marketing:

1. Focus on every opportunity for generating revenues from wireless portal. InfoSpace is considered as a very successful wireless portal. It generates revenues

in three ways, as indicated by its chairman, Naveen Jain (Fletcher, 2000):

(a) When signing up with Infospace, merchants and Internet service providers, are assured of reach millions of wireless customers.
(b) Merchants are required to pay for a percentage (about 2–25%) for each business transaction generated from the wireless portal.
(c) Wireless providers are obliged to pay a per-subscriber, per month fee.

2. Keep brand identity in wireless portal. The service provider is free to negotiate with content providers and merchants who are shown on its portal as long as they agree to carry the brand name in their contents. Currently, Rupp (2002) reported that BT (British Telecom) has adopted this business model, and it seems to be successful in terms of reducing churn rate (the rate at which subscribers stop using the service) and enhancing ARPU (average revenue per user).
3. Charging for contents. "The business model based on paid content is simple and transparent—content providers know what they want and what revenues they will get," said Eden Zoller, the Senior Analyst of Ovum (Fricke, 2001). This is because subscribers are expected to pay for the content, which has perceived value, personalized and suitable for the reception by their mobile devices such as PDAs or Web-enabled mobile phones.

MEASURMENT OF MARKETING PERFORMANCE

Performance-Measurement Management

Korhonen (2001) indicated the following factors for performance-measurement:

1. Traffic levels within the network (traffic load on the radio interface and the utility of resources within the network nodes).
2. Verification of the network configuration (verifications are based on fact findings).
3. Resource-access measurements (at regular time intervals across the network).
4. Quality of service (QoS; attributes as experienced by subscribers).
5. Resource availability (availability of the chosen resources at different phases of the life cycle of the system).

The performance-measurement reports are expected to produce at a particular frequency known as the granularity period. Measurement samples are collected during the granularity period. The results can be kept in the local network entity as files or sent to the concerned parties all at once. The network manager can access these files whenever he or she wants to. Finally, data must be presented to the configuration-management application.

In addition, it is necessary to have a strong system of fraud management for detecting and preventing fraud in mobile telecommunication networks (Korhonen, 2001) to ensure that quality wireless services are provided to customers.

Marketing Effectiveness and Efficiency Measures

Steuernagel (1999, p. 106) defined the effectiveness and efficiency as follows:

1. "Effectiveness is the degree to which the marketing process—advertising, sales channels" converts potential targets into customers. There are many way to measure it, such as
   (a) revenue versus sales expense,
   (b) customer gain versus sales expense,
   (c) revenue versus marketing expense,
   (d) customer gain versus marketing expense,
   (e) store traffic or lead generation versus advertising expense,
   (f) customer satisfaction (e.g., overall satisfaction, monthly statement clarity, missed or error calls over the past 30 days, etc.), and
   (g) customer churn.
2. "Efficiency is a measure of long-term revenue versus the total cost to get and keep the customer—the ratio of unit output to input." It may cover the following measures:
   (a) increase in revenue,
   (b) gross sales per salesperson (vs. peers in the same channel),
   (c) leads and sales generated by promotional campaigns, and
   (d) expected revenue per customer divided by the sales and marketing.

LOOKING AT FUTURE

The following represent future challenges for wireless marketers:

- Customer churn rate,
- Wireless consumer applications,
- Data mining,
- Mergers and acquisition,
- Standards adoption,
- Security concerns,
- Customer relationship marketing,
- Roaming hurdles,
- Billing challenges for GPRS, and
- The rise of the application service provider (ASP), which offers cost savings.

To be successful in the dynamic wireless business, it is necessary for marketers to "work closely with other company departments to form an effective value chain that serves the customer" (Kotler & Armstrong, 2001, p. 677). A value chain involves commitments from various departments of an organization—for example, product design, production, marketing, logistics, and delivery—to

perform value-creating activities. The success of a company's value chain depends on the performance of various departments and how well these activities are coordinated (Kotler & Armstrong, p. 677).

In summary, customers are always the center of all marketing activities in m-commerce. A firm's capabilities and its adaptation for value creation based on customer information are vital to sustaining long-term success. "Customers become data. Then data become customers. Finally, both customers and data migrate to real time" (Lamont, 2001, p. 58). Marketers must carefully examine the best ways to work with their customers; they must know how to use the collected data better than their competitors to create more value for customers.

GLOSSARY

Always-on service A 24-hour service supported by GRPS that allows wireless subscribers access to value-added services anytime and anywhere.

Anonymous Nondisclosure of personally identifiable information

Brand community A group whose members have a particular brand preference and are grouped together to form a brand community. Members of this social group are committed to the brand and are willing to share and exchange information.

Churn rate A measure for the percentage of subscribers who switch from one mobile phone service provider to an alternative provider.

Consumer market The business environment in which individual end users purchase goods or services for their personal or household use.

Customer needs Human needs (e.g., basic needs for food and shelter) that are not conceived by marketers.

Customer loyalty A factor used to describe a consumer's faithfulness in a particular brand or store. A loyal customer is one who has a preference for a particular brand or store without considering the alternatives. Marketers differ on how this loyalty can be measured.

Data mining A method for discovering patterns and meaningful relationships of customers' buying behavior from a data bank

Database marketing A process that involves establishing and updating customer information into a data bank. Marketers use the information provided to decide how and when to contact the consumer for marketing activities, such as a sales promotion.

Direct marketing A method of distribution in which sellers contact potential buyers through direct communication channels (e.g., e-mail) seeking customers' responses (e.g., seeking more information, making a trial purchase).

E-tailing (electronic retailing, e-retailing) A form of direct marketing in which all business transactions are arranged through electronic means such as the Internet and cellular phones.

Lifestyle The ways that a consumer spends his or her time and money.

One-to-one (1:1) marketing Designing and implementing the marketing mix (product, price, distribution, and promotion) for the individual buyer. It relies on extensive information about each purchaser's buying needs and behaviors.

Push promotion A traditional promotional method in business-to-business market. Manufacturers offer discounts or incentives to a wholesaler or retailer to push their products to end users (or final consumers).

Pull promotion Manufacturers focus all their promotional efforts (e.g., promotional discounts or complimentary gifts) to attract end users or final consumers who buy their products. Under this structure, final customers would request the products from retailers. Retailers will order the products from the manufacturers to meet expressed consumer demand.

Segmentation An important concept in marketing. Market segmentation divides the market into small, mutually exclusive groups. Each segment consists of customers with similar needs, characteristics, and buying behavior. Different marketing strategies (i.e., product, price, place [or distribution], and promotion) are required for targeting different segments because their needs are different. Firms can select one or more segments to target based on their strength and resources.

Supplementary channel An additional channel to existing distribution or sales outlets for selling products to customers or end user.

Telemarketing A kind of direct marketing method in which sellers use telephone calls to sell a product or service to an individual customer.

Value chain A management tool to identify methods for generating more customer value.

CROSS REFERENCES

See *Consumer Behavior; Marketing Communication Strategies; Personalization and Customization Technologies; Value Chain Analysis.*

REFERENCES

Advisor.com (2002, April 2). Mobile marketing: Consumers say, "Bring it on!" (Marketing Advisor zone, Doc. No. 09546). Retrieved November 20, 2002, from http://advisor.com

Boone, L. E., & Kurtz D. L. (2001). *Contemporary marketing* (10th ed.). Forth Worth, TX: Harcourt College.

Computer Industry Almanac (2002, March 21). Internet users will top 1 billion in 2005, wireless Internet users will reach 48% in 2005 (press release). Retrieved November 22, 2002, from http://www.c-i-a.com/pr032102.htm

e-Marketer (2003). Mobile penetration in the US and Europe. Retrieved February 13, 2003, from http://www.emarketer.com/news/article.php

Fletcher, J. (2000, August 14). InfoSpace looks to wireless for exponential growth. *eCommerce Business,* p. 14.

Fricke, P. (2001, August 22). Study: charging for content lead to wireless success. *CommWeb.com*. Retrieved October 18, 2002, from http://www.commweb.com/article/COM20010822S0003

Godin, S., & Peppers, D. (1999). *Permission marketing: Turning strangers into friends, and friends into customers.* New York: Simon & Schuster.

Gruen, T., & Ferguson, J. M. (1994). Using membership as a marketing tool: Issues and applications. In N. Jagdish & A. Parvatoyar, *Relationship marketing: Theory, method, and application* (pp. 60–64). Atlanta, GA: Center for Relationship Marketing, Roberto C. Goizueta Business School, Emory University.

International Telecommunication Union (2002a, September). ITU Internet report: Internet for a mobile generation—executive summary. Retrieved October 28, 2002, from http://www.itu.int/osg/spu/publications/sales/mobileinternet/execsumFinal.pdf

International Telecommunications Union (2002b). Hong Kong (China) and Denmark top ITU Mobile/Internet Index (press release). Retrieved February 8, 2003, from http://www.itu.int/newsarchive/press_releases/2002/20.html

Korhonen, J. (2001). *Introduction to 3G mobile communications*. Boston: Artech House.

Kotler, P., & Armstrong, G. (2001). *Principles of marketing* (9th ed.). Upper Saddle River, NJ: Prentice Hall.

Lamont, D. (2001). *Conquering the wireless world—the age of m-commerce*. Oxford, UK: Capstone.

Muniz, A. M., Jr., & O'Guinn, T. C. (2001). Brand community. *Journal of Consumer Research, 27,*412–432.

OECD (1999). *The guidelines for consumer protection in the context of electronic commerce*. Retrieved April 18, 2003, from http://www.oecd.org/pdf/M00000000/M00000363.pdf

Oxford Advanced Learner's English-Chinese Dictionary (4th ed.) (1994). Oxford, UK: Oxford University Press.

Richter, F.-J., & Mar, P. C. M. (2002). Recreating Asia—vision for a new century. In *World Economic Forum—committed to improving the states of the world*. Singapore: Wiley.

Rosenspan, A. (2000, June). Participation marketing. *Direct Marketing,* pp. 54–55.

Rupp, W. T. (2002, July–August). Mobile commerce: New revenue machine or black hole? *Business Horizon*, pp. 26–29.

Steuernagel, R. A. (1999). *Wireless marketing*. New York: Wiley.

Wells, W., Burnett, J., & Moriarty, S. (2003). *Advertising—principles and practice* (6th ed.). Upper Saddle River, NJ: Prentice Hall.

FURTHER READING

Advisor.com (2002). Wireless Internet's future depends on service pricing (Wireless Tech Advisor, Advisor zone, Doc. No. 11487). Retrieved November 20, 2002, from http://advisor.com

International Telecommunications Union (2002c, December 9). Telecom Asia 2002 Show. Special advertising section. *Business Week* (Asian ed.). pp. 36–52

Lamont, D. (1997). Salmon day: The end of the beginning for global business. Oxford, UK: Capstone.

Shama A. (2001, September–October). E-coms and their marketing strategies. Business Horizons, pp. 14–20.

X

XBRL (Extensible Business Reporting Language): Business Reporting with XML

J. Efrim Boritz, *University of Waterloo, Canada*
Won Gyun No, *University of Waterloo, Canada*

INTRODUCTION

Since Pacioli defined the double-entry bookkeeping system in his 1494 book, *Summa de Arithmetica, Geometria, Proportioni et Proportionalita*, there have been many new developments in accounting, and these continue today in response to the demands of business and other organizations, and users. Accounting has contributed to economic prosperity and will continue do so in the future. Indeed, with the explosion of interest in communicating business information over the Internet, the recent initiative to create and implement an XBRL (Extensible Business Reporting Language) promises to dramatically enhance the speed and ease of information exchange for enhanced analysis and decision making. In this chapter, we trace the development of XBRL, from its conceptual origins in SGML and XML, providing examples of its application to financial information, including illustrations of the steps involved in creating an XBRL document. An appendix provides a detailed illustration of the steps involved in creating an XBRL document. We conclude with a summary of the benefits and limitations of XBRL.

XML: A NEW PARADIGM FOR INTERNET DOCUMENTS

Today, most B2B (business-to-business) and B2C (business-to-commerce), and many P2P (person-to-person), interactions involve the exchange of information over the Internet. In fact, the empowerment of information providers to easily and cheaply distribute electronic documents via the Internet has fueled the astonishing growth of the World Wide Web (henceforth, the Web). Today, most documents on the Web are stored and transmitted in HTML (HyperText Markup Language), the Web's lingua franca. HTML is a simple language that was developed to provide hypertext and multimedia functions for the Internet. However, as documents on the Web have grown larger and more complex, information providers have begun to encounter limitations in the functionality of HTML attributable to its lack of extensibility, structure, and data checking. These limitations prevent HTML from being a universal information exchange method. HTML is based on SGML (Standard Generalized Markup Language), the international standard for defining descriptions of the structure and formatting of different types of electronic documents. SGML is complex, difficult, and costly to use. To overcome these limitations and to extend Web technology, XML (Extensible Markup Language) was developed, started by John Bosak in 1996, and established by the W3C (World Wide Web Consortium) as a standard in 1998. W3C was created in October 1994 to lead the World Wide Web to its full potential by developing common protocols that promote its evolution and ensure its interoperability. W3C has around 500 member organizations from all over the world and has earned international recognition for its contributions to the growth of the Web.

In contrast to HTML, which was designed to display data and to focus on its appearance, XML was designed to provide structure and to validate the content of documents. XML not only removes the limitations of HTML, it

also facilitates more precise declarations of content, more effective and efficient information exchange, and more meaningful search results. XML includes self-explanatory data within a document; thus, it can be used for universal information exchange over the Internet. This section outlines the similarities and differences between SGML, HTML, and XML, concluding with a summary of the benefits of XML and setting the stage for the subsequent discussion of XBRL.

SGML (Standard Generalized Markup Language)

GML (General Markup Language) was developed at IBM in 1969 by Charles F. Goldfarb, Ed Mosher, and Ray Lorie. Markup refers to the sequence of characters or other symbols that are inserted at certain places in a text or word processing file to indicate how the file should look when it is printed or displayed, or to describe the document's logical structure. The markup indicators are often called *tags*. GML was not merely an alternative to procedural markup but the logical representation that motivated all processing. Publishing companies implemented it because they needed a means of tagging the contents of a document so that text could be presented in a number of different ways. This approach combined two traditions, one then about 25 years old and the other around 500 years old. (A third tradition, the computer programmer's strategy of tying markup to specific interpreters, which went back to the first versions of ROFF at MIT in about 1962, was intentionally set aside as violating the independence of structural information from presentation processing.) The 500-year-old printing and publishing industry tradition started when the first editor needed to give unambiguous instructions to more than one typesetter and developed his own markup language to do it. According to Smith (1996, pp. 75–92), this goes back at least to Anton Koberger's printing of the Wurzburg Pslater in 1486.

Most of these instructions had the common characteristic of using instructional tags in a format, such as <start bold>some text<end bold>, to communicate display formatting instructions to typesetters and other artisans as clearly and unambiguously as possible. The 25-year-old computing tradition was reflected in early text processing applications, such as DIALOG and COLEX, developed in the 1960s. These applications generally tagged data by type at the time of data entry to make it easy to apply Boolean logic in text searches on files prepared using tape-to-tape sorts. Thus, a file entry would contain both data and labeling information about the meaning and role of that data, such as in the following example:

```
PUBDATE: JULY 26, 1959
LIB: 105DWC/PEMBROOK
PUB: RS, NY
AUTH: JOHN SCARNOUGH, BRANSTON GRECHI
TITLE: LOAD CONDITIONS FOR POLYHEDRAL RISERS
ABSTRACT: REVIEW OF STRUCTURAL INTEGRITY
  FAILURE CONDITIONS FOR MECHANICAL
  RISERS IN INTERVAL SUPPORT STRUCTURES.
```

In the mid-1980s, SGML was established by the International Organization for Standardization (ISO 8879:1986) as an international standard for defining and using document structure and content. ISO, founded in 1947, is a worldwide federation of national standards bodies from some 100 countries, one from each country. Among the standards it fosters is OSI (Open Systems Interconnection), a universal reference model for communication protocols. Many countries have national standards organizations, such as the American National Standards Institute (ANSI), that participate in and contribute to ISO standards. SGML incorporates both data labeling and data presentation information but leaves procedural issues entirely to the rendering application.

Because SGML is a generalized theoretical specification, actual use requires selection of a specific DTD (Document Type Definition). A DTD defines the tags the document type will use, what they mean, and whether, and if so to what extent, individual tags can be nested. For example, HTML is a SGML DTD. SGML also requires use of an application that will correctly interpret the DTD in combination with the document text to output either data for use by another application or instructions for a rendering engine. SGML also requires a data processing application or a rendering application that will output the document on a specific display device, such as a screen or printer. Web browsers like Netscape or Internet Explorer contain a rendering tool, such as a compiler or document handler, that combines text with HTML markup to create displayed pages using a set of internal rules known as style sheets. For example <TITLE>text</TITLE> is interpreted as labeling information giving the document title as "text"; <P> as an instruction to begin a left-justified paragraph within the current text block, and <H1>text</H1> as an instruction to display "text" on a line (or lines) by itself using the font type and color set in the enclosing block but at about 3.1 times the user's global default font size.

SGML was intended to be a language that would account for every possible document format and presentation. Thus, it enables users to create tags, to define document formats, and to exchange data among various applications. Because SGML is system and platform independent and can save and validate a document's structure, it can be used in various ways, such as for searching and exchanging data. However, SGML is complex and contains many optional features that are not needed by many Internet applications. Furthermore, it is costly to develop software that supports SGML. As a result, there are few SGML applications for the Internet. Two examples are Xmetal and Wordperfect. Figure 1 shows an SGML document that describes customer e-mail information.

HTML (HyperText Markup Language)

As mentioned earlier, HTML, the basic language for creating a Web page, is based on SGML. HTML consists of a set of markup symbols inserted in a file intended for display on a Web browser. The markup tags tell the Web browser how to display a Web page's words and images for the user. Each individual markup tag is referred to as an element. HTML uses predefined tags, and the meaning of these tags is well understood; for example, means a paragraph, means end of a paragraph, and <table> means a table. Thus, the text "Current Assets, Cash

Figure 1: SGML example.

and Cash equivalents, $12,345" can be marked up using HTML as

```
<table border="1" cellpadding="0"
  cellspacing="0" width="70%" align="Center">
  <tr>
     <td width="70%"><p align="left">Current
       Assets</p></td>
     <td width="30%"></td>
  </tr>
  <tr>
     <td width="70%"><p align="left">Cash
       and Cash Equivalents</p></td>
     <td width="30%"><p align="center">
       $12,345</p></td>
  </tr>
</table>
```

This markup produces the table below when rendered on a display using the postscript language. However, the result can be unpredictably different when rendered using another combination of Web browser, graphics processor, desktop settings, and screen.

Current Assets	
Cash and cash equivalents	$12,345

HTML also adds hyperlink functions to simple documents using tags. A hyperlink function enables a user to jump from document location to document location within the same document or documents at physically distant locations connected by the Internet. Hypertext is the organization of information units into connected associations that a user can choose to make. An instance of such an association is called a link or hypertext link. Hypertext was the main concept that led to the invention of the Web, which consists of information content connected by hypertext links.

HTML is easy to learn and use. Its simplicity and convenience have aided the explosion of interest in the Internet. Figure 2 contains an example of an HTML document and Figure 3 shows the HTML example from Figure 2 as it would appear in a Web browser such as Internet Explorer.

As mentioned previously, HTML has some fundamental limitations. It merely facilitates access to text and multimedia. It does not allow intelligent search, data exchange, and non-HTML forms such as spreadsheets and databases. Although HTML tags generally indicate only how the content should appear, XML tags indicate what the content is. For example, a financial statement prepared using HTML displays the financial statement itself but cannot communicate information about the classification of numbers within categories or subtotals like current assets or cash and cash equivalents. In contrast, XML can give information about the meaning of the numbers in the financial statement as well as displaying them. Content, therefore, becomes more readily accessible through XML. As e-commerce grows, it becomes important to exchange data, use more meaningful search, manipulate data, and generate multiple views of the data. As a result, HTML's limitations and XML's virtues are becoming increasingly noticeable.

XML Document

XML stands for Extensible Markup Language. It is extensible because the language can be extended by anyone who wants to create additional tags for new and unforeseen purposes. It is a markup language because XML is a method of tagging information using accepted rules and formats to give definition to text and symbols. XML was invented by adopting the key functions of SGML while excluding the less essential ones. In fact, SGML can be used without modification and can be converted to XML. (Further information may be found at http://www.w3.org/TR/NOTE-sgml-xml-971215.html, Clark, n.d.) Furthermore, existing HTML can continue to be used, but more complicated and more highly structured documents can be created using XML.

The basic structure of XML is similar to HTML in many respects. XML documents consist of XML elements. Basically, these elements involve a start tag such as <TITLE>, an end tag such as </TITLE>, and the

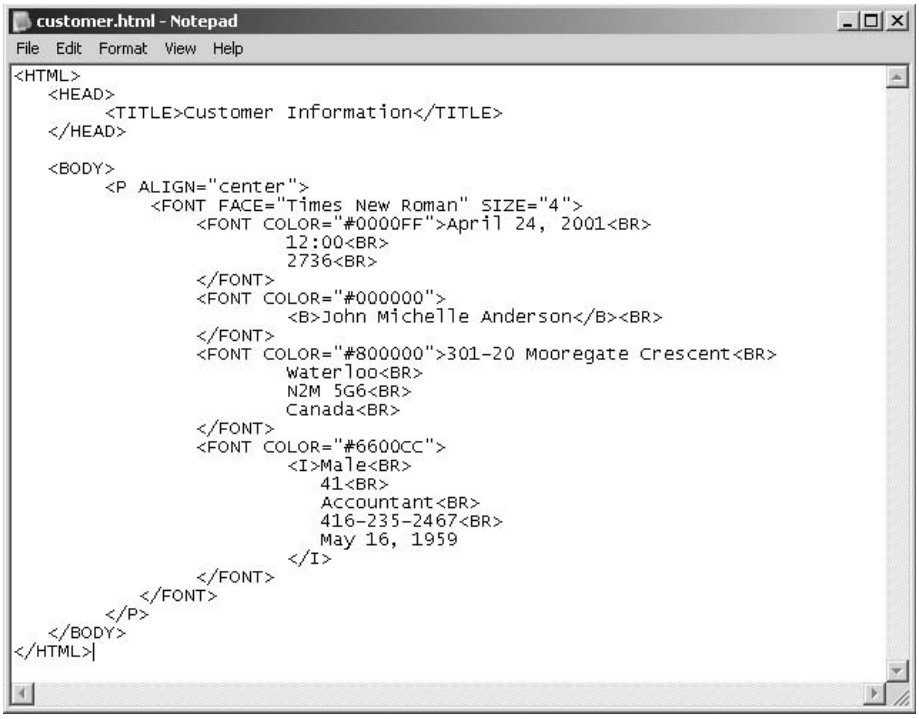


Figure 2: HTML example.

content between the two tags. Figure 4 shows an XML document that contains customer information. The code for this example and all the other examples and taxonomies described in this paper is available in full online at http://arts.uwaterloo.ca/~jeboritz/XBRL/

Unlike HTML, XML tags indicate what each item of data means; for example, tags such as <DATE>, <ADDRESS>, and <PROFILE> convey meaning. In the example in Figure 4, the data defined by the <PROFILE> tag indicates that the data represents customer profile.

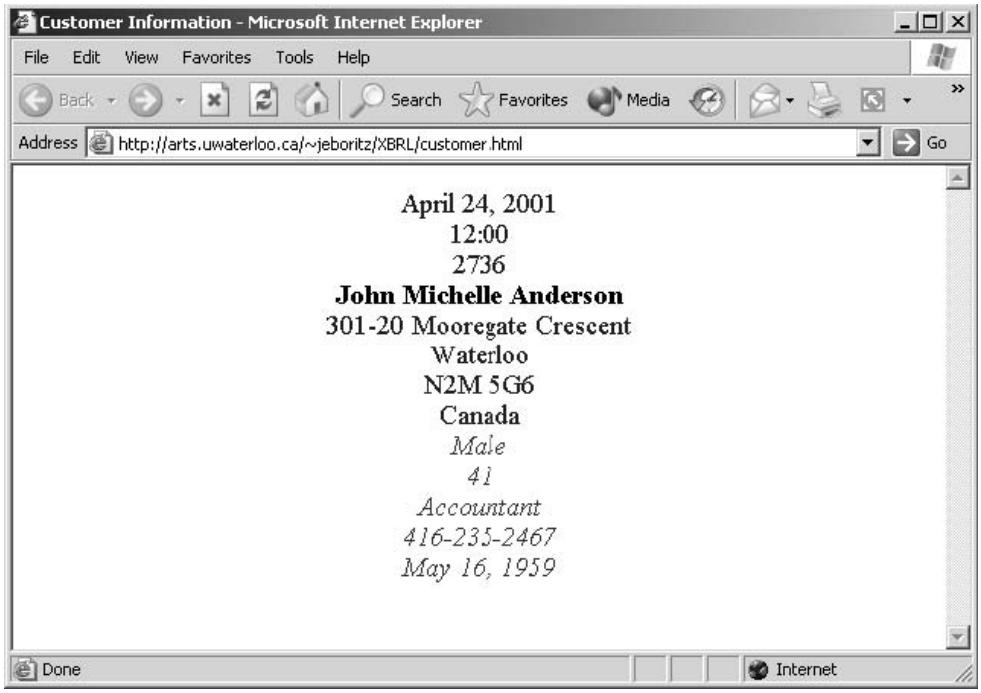


Figure 3: HTML example in Internet Explorer.

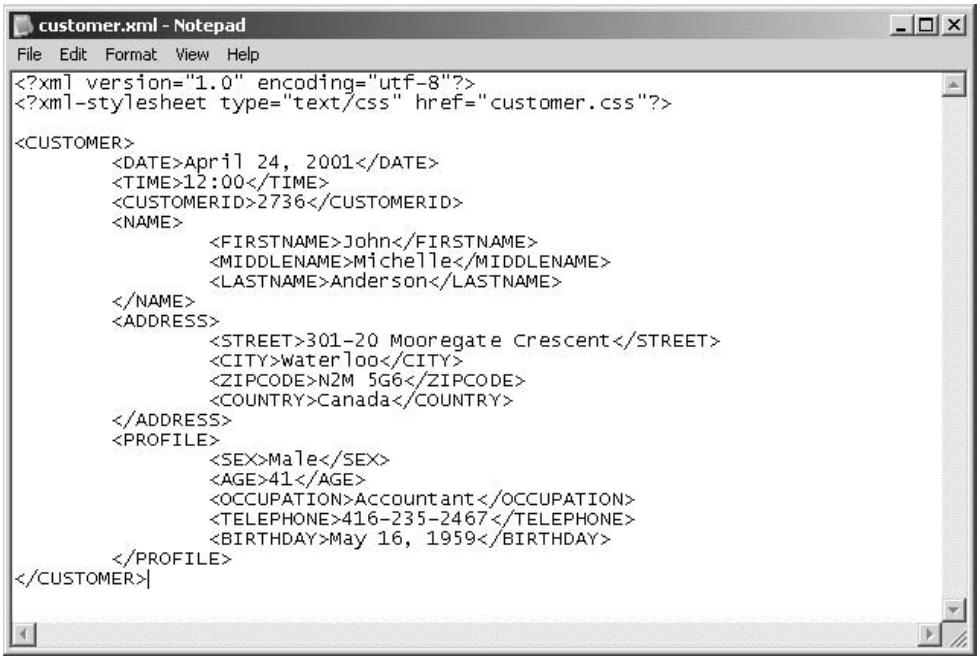


Figure 4: XML example.

Therefore, recipients of this document can decode the XML data and use it for their own purposes. For instance, a customer manager might use it to find customers who live in the Waterloo area (<CITY>Waterloo</CITY>).

Because HTML tags are predefined and understood by Web browsers, the Web browsers can display HTML-tagged documents. In contrast, because XML enables users to create any tags they need, the meaning of these tags will not necessarily be understood by a Web browser unless they are the preexisting HTML commands. There is no way for a generic Web browser to anticipate all possible tags and contain all the necessary rules for displaying them. Thus, to display XML documents in a Web browser, it is necessary to have a mechanism to describe how the document should be displayed. This is done by means of a style sheet. A style sheet is prepared using style sheet language. Two of the most popular style sheet languages are CSS (Cascading Style Sheets) and XSLT (Extensible Stylesheet Language Transformations). CSS provides procedural control by putting the additional presentation control information needed to force a browser to override its internal style sheet inside the HTML file. Thus, the following command overrides the browser's internal style sheet to turn text between <H1> and </H1> blue while forcing it to a 12-point font size regardless of user settings.

```
<STYLE TYPE=text/css>
  H1 {color:blue font-size:12.000000pt}
</STYLE>
```

When a browser reads a style sheet, it will format the document accordingly. There are three ways of linking a style sheet to HTML: external style sheet, internal style sheet, and inline styles. An external style sheet can be created with any text editor, and a CSS style sheet should be saved with a "css" extension. Figure 5 shows an example of an external style sheet. An internal style sheet should be used when a document has a unique style. The internal style sheet is defined in the head section by using the <style> tag. The example shows an internal style sheet. Finally, inline styles can be used with the style attribute in the relevant tag. The style attribute can contain any CSS property. (Further information may be found at http://www.htmlhelp.com/reference/css/style-html.html, Web Design Group, n.d.)

Figure 5 contains the CSS code used to represent the XML example in Figure 4, so that it can be displayed by a Web browser exactly like Figure 3. There are several tools available to create the CSS code (e.g., Microsoft FrontPage).

After a XML document and a style sheet for that document are prepared, the XML document can be shown in the Web browser by including an instruction in the XML document specifying the style sheet to be used, as in Figure 4: <?xml-stylesheet type = "text/css" href = "customer.css"?>. Of course, an HTML document is but one way of presenting an XML document. Because XML separates content from presentation format, through the use of style sheets and other programmatic methods, the content in XML can be presented in several ways, such as an HTML document, text document, and spreadsheet.

Benefits of XML

XML enhances the power and flexibility of Web-based applications and other business software packages. It is an open standard and is system and platform independent. Also, it is free—the XML specifications of tags and attributes are developed by the W3C, which is a consortium led by not-for-profit entities. Thus, XML creates a universal way for both formatting and presenting data and enables putting structured data

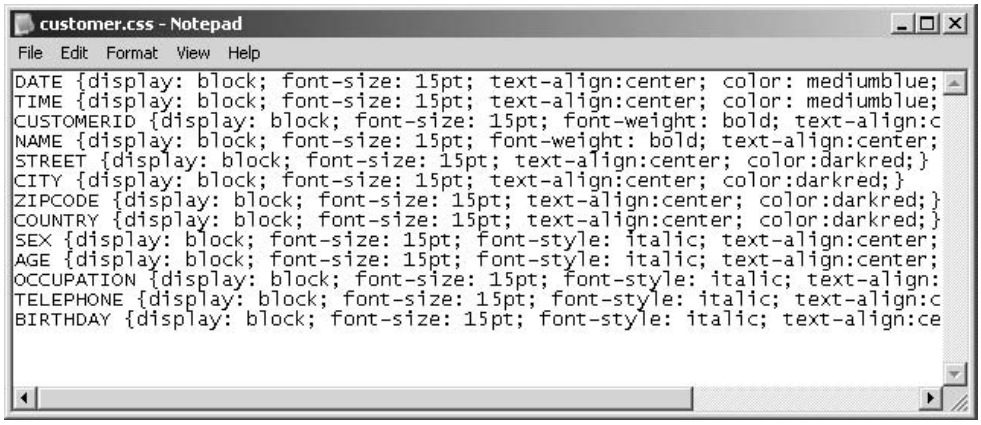


Figure 5: CSS example.

in a text file. Because data in XML are coded with tags that describe content and structure, the data presented in a XML format can be parsed, edited, and manipulated. Searches can produce more accurate and relevant outputs. Data can be exchanged and processed without modification on any software and any hardware platform because XML-based data are self-describing. These capabilities have potential application to B2B communications, transaction processing, and data transfers between various systems and platforms. Once an XML file has been delivered to users, they can view it in different ways. Because XML is extensible, it also allows users to create their own validation tools, including DTDs and XML schema, effectively creating extensible tag sets that can be used for multiple applications.

As mentioned previously, a DTD is a specific definition that follows the rules of the SGML. It defines elements, element attributes, and values and describes specifications about which elements can be contained in others. A DTD accompanies a document and can be used as a validation tool.

An XML schema is an XML-based alternative to a DTD. It is developed to provide XML with flexibilities that a DTD does not possess to meet users' needs. For example, with a DTD it is difficult to validate the correctness of data, to work with data from a database, and to describe permissible document content because a DTD does not have support for data types. In addition, a DTD is not an XML document; therefore, users have to learn another language. In contrast, an XML schema describes the structure, content, and semantics of an XML document. Thus, an XML schema provides a means for expressing shared vocabularies and allows machines to carry out rules made by developers. (Further information may be found at the XML schema Web site, http://www.w3.org/XML/Schema.html, W3C, n.d.b.) Also, users do not have to learn another language to create an XML schema because it is written in XML.

XML is being extended with several additional standards that add styles (XSLT), linking (XLink), and referencing ability to the core XML set of capabilities. XML linking language (XLink) is a method of creating and describing hyperlinks that support both traditional HTML and extended links, which provide more functionality than traditional HTML links. (Further information may be found at the XLink Web site, http://www.w3.org/TR/xlink/, W3C, n.d.a.)

XBRL (EXTENSIBLE BUSINESS REPORTING LANGUAGE)

XBRL (Extensible Business Reporting Language), formerly code-named XFRML (XML-based Financial Reporting Markup Language), is the financial profession's adaptation of XML for financial reporting. A joint industry and government consortium was established for this purpose in the fall of 1999, including the American Institute of Certified Public Accountants (AICPA), six information technology companies, and the five largest accounting and professional services firms. (Information on XBRL member organizations is available at the XBRL Web site, http://www.xbrl.org/.) The consortium developed an XML-based specification for the preparation and exchange of financial reports and data. This freely available and open specification provides a method by which financial professionals can prepare, extract, analyze, and exchange business reports, like financial statements, and the information they contain.

The main objective of financial reporting is to provide useful information to users for their decision making purposes. By providing financial information to intranet, extranet, and corporate Web sites, entities can help users obtain more easily and on a more timely basis the information they need. However, because there are no common, generally accepted formats for describing business reporting data, it is difficult to generate reporting formats tailored to different users' needs and to exchange data among applications. Thus, users seeking to work with data posted on Web sites must reenter or cut and paste the data into their documents or spreadsheets. This is clearly inefficient.

XBRL was created to help address these issues by creating a set of tags recognizable to XML-enabled Web browsers or other applications, such as spreadsheet and database software. Using XBRL, tags are attached to all financial statement data to identify them as asset, current asset, liability, capital, profit, and so forth. Therefore, information users can use a Web browser to visit

companies' Web sites, find the data with the tags (e.g., CashAndCashEquivalents), extract the data, and analyze the data with analytical applications.

XBRL is different from other XML-based specifications, such as FpML, FIX, FinXml, OFX, ebXML, and XML/EDI.

- FpML (Financial Products Markup Language) is an XML-based industry-standard protocol for swaps, derivatives and structured products. (Further information may be found at the FpML Web site, http://www.fpml.org, FpML, n.d.)
- FIX (Financial Information Eschange) defines specific kinds of electronic messages for communicating securities transactions between two parties. (Further information may be found at the FIX Web site, http://www.fixprotocol.org, FIX, n.d.)
- FinXML is an XML-based standard for financial institutions to exchange financial data and to communicate the details of highly structured financial transactions. (Further information may be found at the FinXML Web site, http://www.finxml.org, FinXML, n.d.)
- OFX (Open Financial Exchange) is a specification for the electronic exchange of financial data between financial institutions, businesses, and consumers via the Internet. (Further information may be found at the OFX Web site, http://www.ofx.net, Open Financial Exchange, n.d.)
- ebXML (Electronic Business using XML) is a suite of specifications that enables enterprises to conduct business over the Internet. ebXML provides a standard method for exchanging business messages, conducting trading relationships, communicating data in common terms, and defining and registering business processes. (Further information may be found at the ebXML Web site, http://www.ebxml.org, ebXML, n.d.)
- XML/EDI (Extensible Markup Language/Electronic Data Interchange) provides a standard framework to exchange various types of data—for example an invoice, healthcare claim, or project status—can be searched, decoded, manipulated, and displayed consistently and correctly by first implementing EDI dictionaries and extending vocabulary via online repositories to include business language, rules, and objects. This framework is intended to apply to a variety of information sources: transactions, exchanges via an API (Application Program Interface), Web automation, database portal, catalog, workflow document, or message. (Further information may be found at the XML/EDI web site, http://www.xmledi-group.org, XMLEDI, n.d.)

In contrast with these transaction-oriented specifications, XBRL is reporting oriented. As such, XBRL enables individual investors and financial professionals to search through and extract data from financial statements, then place them in their own applications, simplifying one of the key phases of financial statement analysis. XBRL does not establish new accounting standards (although XBRL.org has a process for approving taxonomies) but is intended to enhance the value or usability of existing standards. Also, it does not require providing additional financial information to outside users.

How XBRL Works

Figure 6 depicts how XBRL would be used. Suppose that a public company, Toronto Inc., wishes to provide financial statements to analysts.

After the company prepares its financial information using its internal accounting system, an XBRL document is created by mapping the financial information to

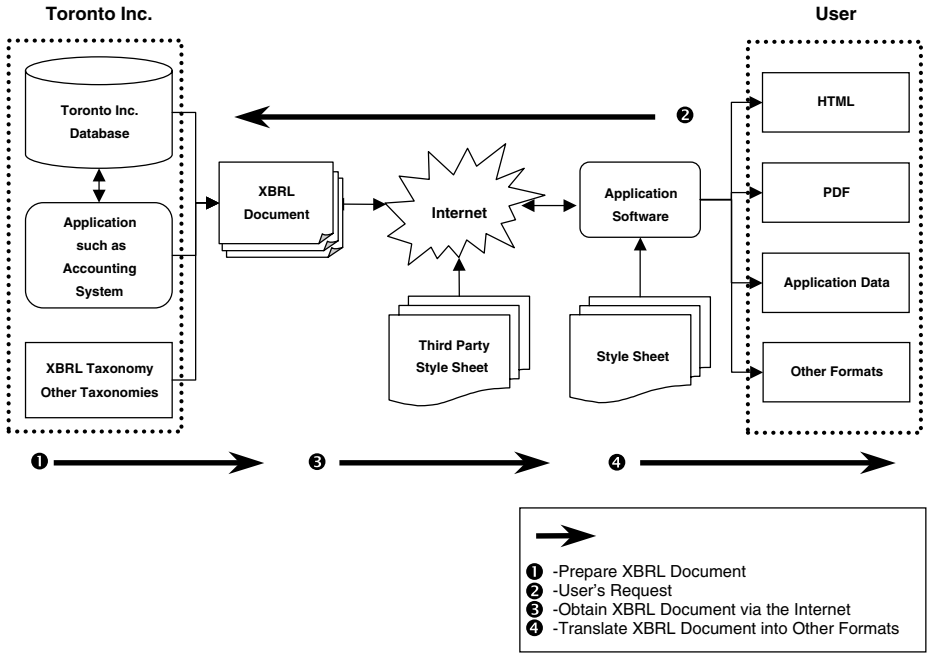


Figure 6: How XBRL works.

XBRL taxonomy elements. A number of new software packages can do this automatically. The created XBRL document is automatically checked to ensure it is proper XBRL. There are several XBRL document validation programs. One is available from XBRL Solutions Inc., at http://www.xbrlsolutions.com. Then, the validated XBRL document is placed on the company's Web site or FTP server.

When users need the information contained in the XBRL document for their analysis, they obtain it on the Internet. Users use the XBRL document for their analysis. If they want to translate the document into HTML, a spreadsheet, or database, they can do so with appropriate style sheets developed by them or by outside software developers. Of course, an XBRL document would not necessarily be viewed by a person in its raw form. XML-enabled software could automatically parse and transform the content of the XBRL document and then transfer it to another system for further processing.

XBRL Taxonomy and XBRL Instance Document

A XBRL document is created by mapping financial information to an XBRL taxonomy that describes financial "facts" and the relationships between them. A taxonomy is a dictionary of the financial terms used in preparing financial statements or other business reports and the corresponding XBRL tags. A XBRL taxonomy can be regarded as an extension of an XML schema that defines elements corresponding to concepts that can be referenced in XBRL documents; for example, the element with the name "nonCurrentAssets.propertyPlantAndEquipmentNet" represents such a concept. A XBRL document has a hierarchical structure that is defined by the taxonomy. Figure 7 contains a graphical illustration of the hierarchical structure of an XBRL document and a taxonomy as seen through a "taxonomy viewer."

A XBRL instance document is an XML document containing XBRL elements. In other words, a company's financial statements, created by using XBRL, are an instance document in which various XBRL elements are embedded based on a specific taxonomy. A common taxonomy enables users to compare several firms' financial statements (assuming they use the same accounting guidelines). Because the same tags are used by all "publishers" of XBRL documents, who rely on the same taxonomy, all users of those documents will recognize the tagged data the same way. For example, the tag <group

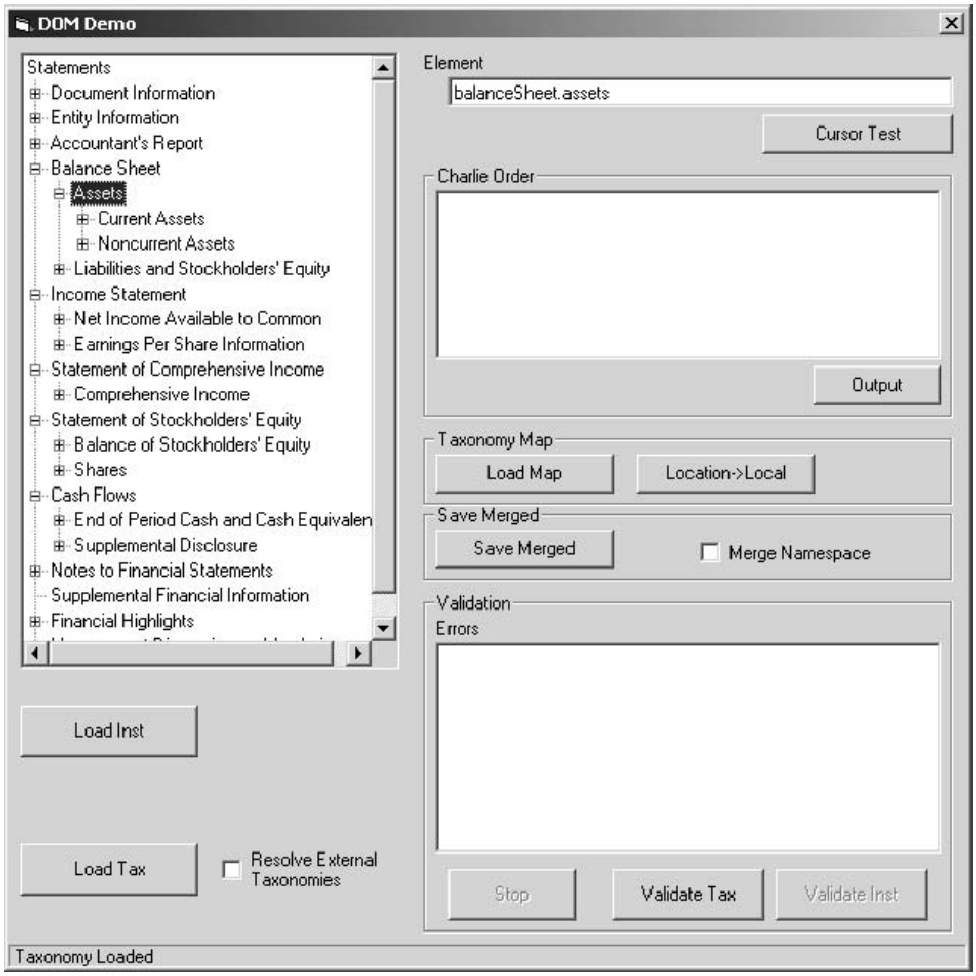


Figure 7: XBRL taxonomy as seen through a taxonomy viewer.

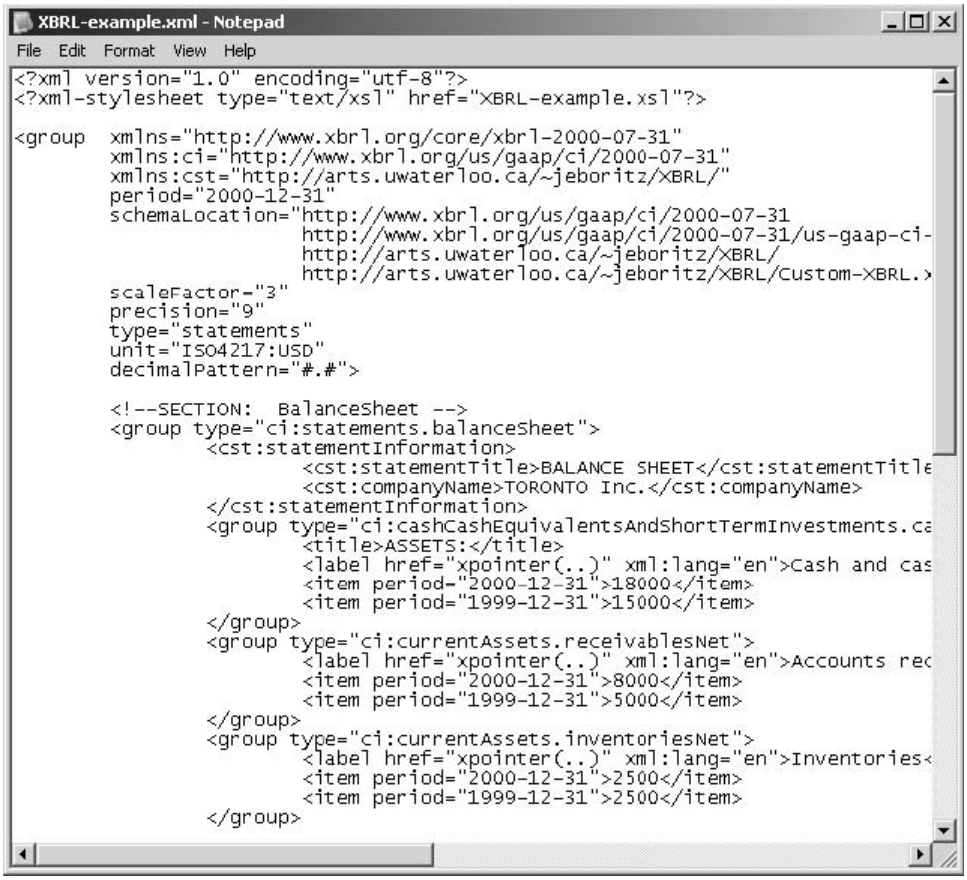


Figure 8: XBRL example.

type = "ci:balanceSheet.assets"> indicates that the data after the tag relates to assets on the balance sheet. Figure 8 shows an example of an XBRL instance document. (Please note that all the examples used in this chapter are based on version 1.0 of XBRL. The XBRL specification version 2.0 was publicly announced on Dec. 14, 2001. This specification is to be used to create XBRL taxonomies. However, at the time this chapter was prepared no XBRL taxonomy based on version 2.0 was available. Version 2.0 is significantly different from version 1.0, but the key concepts described in this chapter are still applicable.)

The sample XBRL document in Figure 8 is based on the XBRL taxonomy for business reporting of commercial and industrial companies, under U.S. Generally Accepted Accounting Principles (US GAAP), dated 2000-07-31. Figure 9 shows a part of this taxonomy. (A new version of this taxonomy was nearing completion at the time this chapter was written. The new version will have significant differences from the version illustrated in this chapter; however, the key concepts discussed here still apply.)

A XBRL taxonomy has several advantages. First, it makes it possible to bring the semantics of financial information to the surface, making it more transparent. Second, an XBRL taxonomy can make it easier to analyze how similar companies applied various accounting rules or methods. Third, an XBRL taxonomy provides contextual information that can enhance information search using XBRL-based search engines to obtain more complete and more accurate search results. Finally, an XBRL taxonomy can be used for converting an XBRL document based on one country's generally accepted accounting principles (GAAP) to an XBRL document based on another country's GAAP. For example, an XBRL document prepared by using a U.S. taxonomy can be converted to an XBRL document based on International Accounting Standards (IAS) taxonomy. This may become an important function because hundreds of national and industry taxonomies are being developed.

At present, several taxonomies have been developed such as US GAAP CI, the XBRL taxonomy for business reporting of commercial and industrial companies under US GAAP, the XBRL GL (General Ledger) taxonomy, and the IAS GAAP CI taxonomy. Also, several national taxonomies are under development: Canadian GAAP CI, German GAAP CI, Australian GAAP CI, New Zealand GAAP CI, and Singapore GAAP CI. Also, several prominent organizations have announced commitments to XBRL, including the Australian Prudential Regulatory Authority, UK Inland Revenue, U.S. FDIC, NASDAQ, Microsoft, Bank of America, Morgan Stanley, and Reuters. Information about XBRL taxonomies may be found at http://www.xbrl.org, XBRL Home Page (2000). Information about the Australian Prudential Regulatory Authority may be found at http://www.apra.gov.au, APRA (n.d.). Information related to Microsoft's commitment to XBRL may be found at Microsoft's web site, http://www.microsoft.com/msft, Microsoft (n.d.). Information on companies and organizations that

Figure 9: Taxonomy for business reporting of commercial and industrial companies, US GAAP.

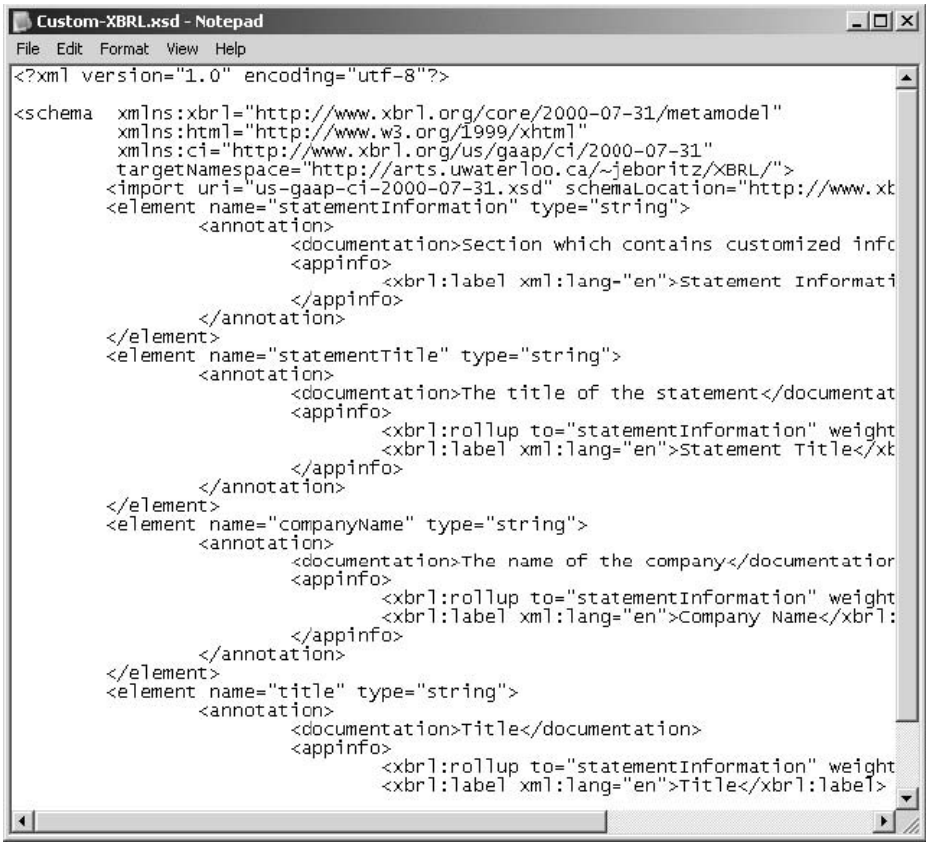


Figure 10: Customized taxonomy example.

provide XBRL statements may be found at the XBRL express web site, http://www.edgar-online.com/xbrl.

Extensibility

Because XBRL is an application of XML, it is extensible. Therefore, if a taxonomy does not contain tags that meet the users' needs, the users can create their own tags. An example of a customized taxonomy is shown in Figure 10. The customized taxonomy contains added elements to describe statement information, statement title, and company name.

Style Sheets

Although XBRL documents can be easily handled by software applications, they are not easily readable by people. However, XBRL documents can be transformed into user-understandable formats, such Web pages, text documents, and other XBRL documents, with the help of style sheets. As mentioned previously, Web browsers such as Internet Explorer do not have built-in semantics that enable them to process a labeling tag such as <Assets.CurrentAssets>. Thus, for an XBRL document to be displayed by a browser, it must first be transformed into a document the browser can render.

As mentioned previously, CSS (Cascading Style Sheets) and XSLT (Extensible Stylesheet Language Transformations) are the two most popular style sheet languages. CSS and XSLT overlap to some extent. CSS is more broadly supported than XSLT. Most Web browsers support CSS, but only a few accept XSLT. Thus, it is likely most of CSS (CSS1 and some of CSS2) to be well supported by popular Web browsers, such as Netscape and Internet Explorer. CSS1 (Cascading Style Sheets level 1) is a W3C recommendation. It describes the CSS language as well as a basic formatting model. CSS 2 (Cascading Style Sheets level 2), which is also a W3C recommendation, builds on CSS1. It includes media-specific style sheets (e.g., printers and aural devices), and element positioning and tables. Although many HTML users and developers are familiar with CSS, it only provides for formatting of contents. It does not allow users to change or reorder contents. (Further information may be found at http://www.w3c.org/Style/CSS, W3C, n.d.c.)

Although XSLT is more complicated than CSS and is not well supported, it is a more powerful and flexible style sheet language than CSS. XSLT is currently the only style sheet language designed specifically for use with XML. XSLT can transform XML into other documents, such as HTML or database, filter and sort XML data, and format XML data. XSLT has evolved from the early XSL standard. XSL consists of a language for transforming XML documents and a language for formatting XML documents. The XSL formatting language, often called

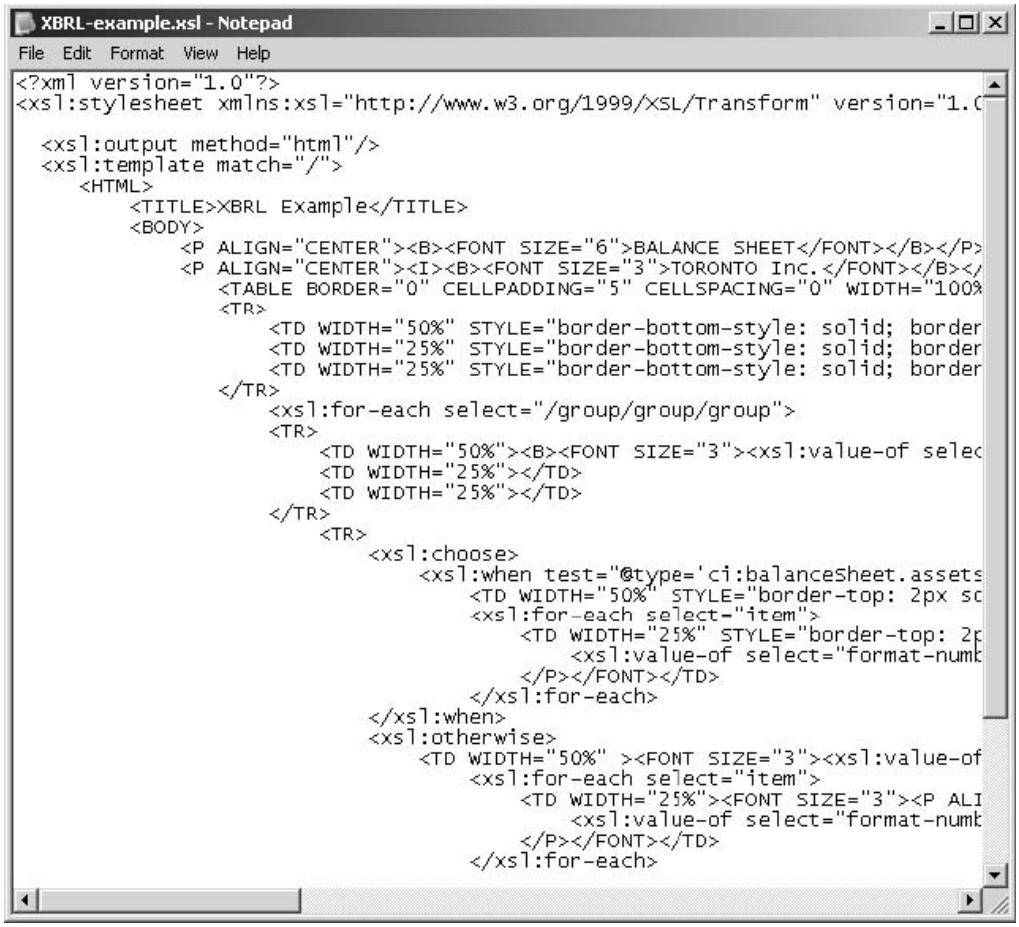


Figure 11: XSLT example.

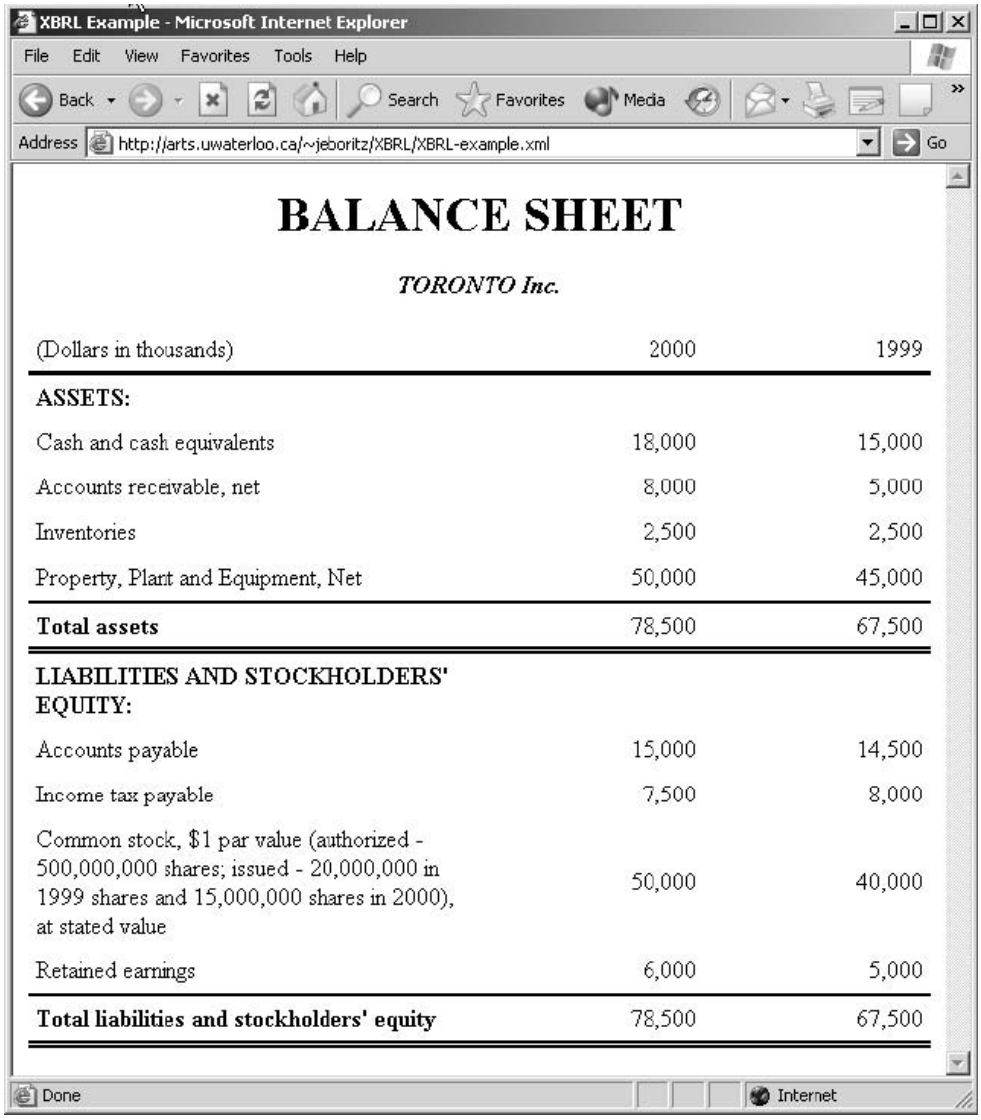


Figure 12: XBRL example in Internet Explorer.

XSL-FO (Extensible Stylesheet Language Formatting Objects), provides means to display data in some format and/or media. XSL transformations language, known as XSLT, provides a means of parsing an XML documents into a tree of nodes, and then converting the source tree into a result tree. XSLT was proposed and later accepted as a separate standard for XML data transformation only. XSL is now generally referred to as XSL-FO (XSL Formatting Objects), to distinguish it from XSLT. XSLT can transform selected XML elements into HTML elements. In addition, XSLT can add new elements into the output file, or it can remove elements. It can also rearrange and sort the elements. Figure 11 shows an XSLT document that is applied to the XBRL example in Figure 8. Further information may be found at the XSL Web site, http://www.w3.org/TR/xsl, W3C, n.d.d.)

A XBRL document can be shown in a Web browser by including an instruction in the XBRL document to specify the XSLT style sheet to be used (<?xml-stylesheet type = "text/xsl" href = "XBRL-example.xsl"?>). To simplify the example, the style sheet is applied to the XML document by embedding the stylesheet command in the document. However, it should be noted that embedding the stylesheet command in an XML document is only one way of displaying an XML document. Another method of handling an XML document is to use XSLT to transform an XBRL or XML document into other formats, such as HTML, text, spreadsheet, and database formats. Figure 12 shows the XBRL example in Internet Explorer.

Users might want to move data in and out of special applications, such as database and spreadsheet software. For example, users might want to view, update, and review the data in a spreadsheet as an intermediate step in the context of a larger business process. Currently, most major office suite software supports the storage and manipulation of XML documents. Microsoft Office XP, for instance, supports XML document files. Therefore, with XML support built into Excel, users can load data

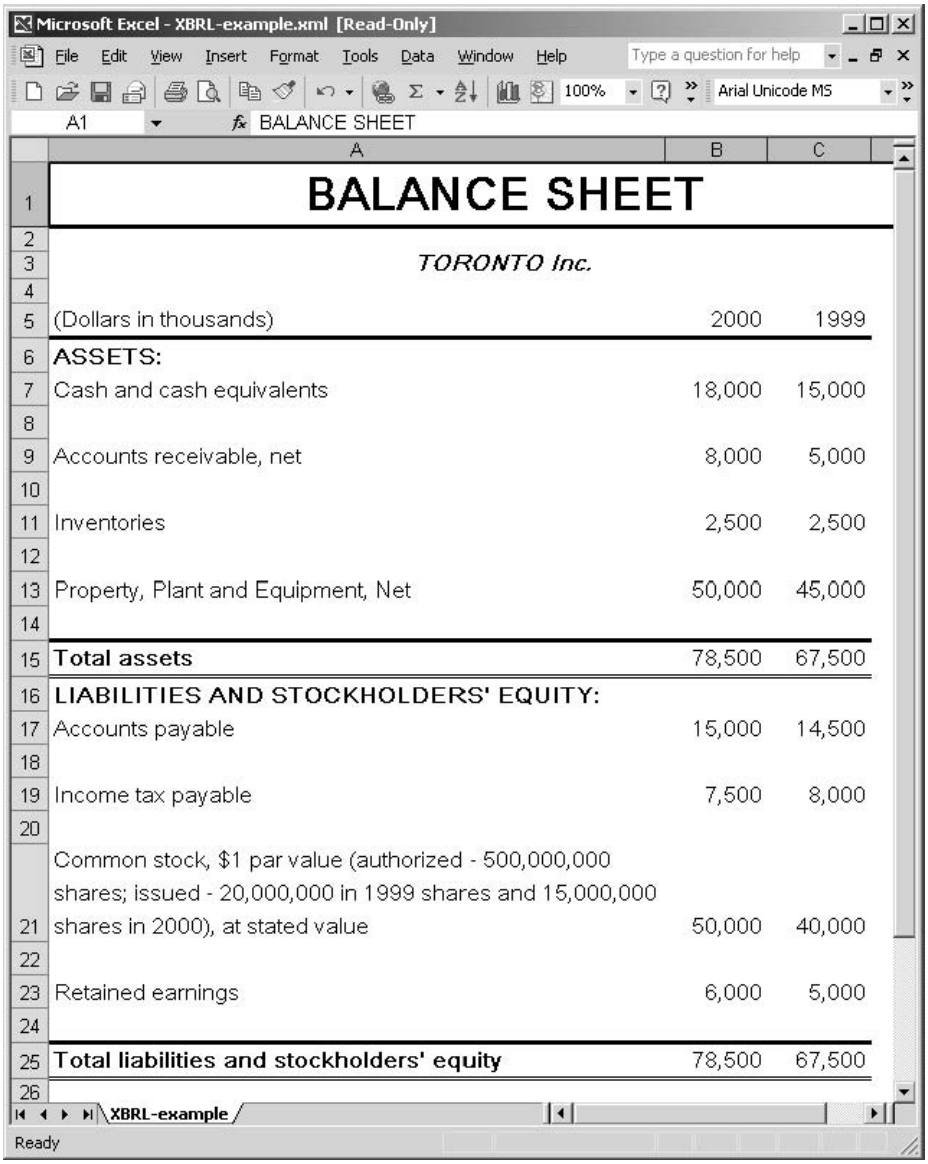


Figure 13: XBRL example in Excel.

from an XBRL document into Excel and apply their own analytic procedures on the data. Figure 13 shows an XBRL example document that is loaded into Microsoft Excel.

Appendix I provides a detailed illustrative example of the steps involved in creating an XBRL document.

Appendix II lists additional sources of information on XML and XBRL.

SUMMARY OF ADVANTAGES AND LIMITATIONS OF XBRL

Advantages of XBRL

XBRL provides a number of important benefits. It is technology independent; thus, XBRL is relevant for all users of financial information, regardless of system or platform used. Second, XBRL tags describe content and structure. Therefore, by searching tagged information, users can obtain more reliable results more efficiently. Third, because XBRL documents are prepared using a taxonomy, data can be exchanged and processed without modification. This permits interchangeability of data and cuts down on data manipulation costs. Fourth, XBRL documents can be transformed to several formats, such as a Web page and a data file for spreadsheet and database software. Consequently, data in XBRL may be displayed in a Web browser, sent to a database, sent to a printer, and used to create other XBRL documents. Fifth, because XBRL facilitates paperless financial reporting, the cost of producing financial and regulatory information can be substantially reduced. Finally, XBRL enhances the analysis of multiple company financial information; users can obtain and analyze several companies' financial data simultaneously (assuming they follow the same generally accepted accounting principles).

Limitations of XBRL

Although XBRL provides a number of important benefits, it also has limitations. First, it is important to note that XBRL provides tags for conventional financial statements and does not address other methods of displaying financial data, such as formulas, graphs, and charts. In fact, XBRL is not primarily designed for easy rendering of information for human consumption—its primary objective is to enhance efficiency of data transfer and archiving. Many accounting professionals and users have suggested that current financial statements do not fully live up to users' expectations because the conventional financial statement reporting format may not meet the needs of users who prefer various types of multimedia formats or multidimensional numerical and graphical presentations. XBRL does not provide a standardized format for such displays, focusing instead on conventional financial statement content. Second, XBRL does not currently address the quality of information, for example whether data coded in XBRL are reliable. A number of XBRL taxonomies are being developed, raising the possibility of errors in selecting and applying XBRL taxonomies and codes/tags. Also, information on the Web can be easily created and revised, and its source can be disguised, raising questions about the trustworthiness of information disseminated via the Internet, including information in XBRL format. Third, XBRL currently views the accountant's report as part and parcel of the financial statement package as a whole. It does not provide for an assurance report on an individual financial statement or an individual item in a financial statement, or on the internal controls underlying the financial reports. We deal with these issues in some depth in Boritz and No (2002).

CONCLUSION

Many companies are attempting to disseminate financial information over the Internet. They have set up Intranets, connected the Intranets to the Internet, and have created corporate Web sites to provide employees, investors, financial analysts, and other users with the information they need on a timely basis. However, data must often be reentered or cut and pasted by users seeking to analyze it because there are no common, generally accepted formats for describing business reporting data.

XBRL was created to help address these problems. XBRL is a markup language for documents containing structured financial information. It provides a standardized method to prepare, publish, and exchange financial reports and the information they contain without modification. Thus, XBRL offers technology independence, full interoperability, efficient preparation of financial statements, and efficient extraction of financial information for analysis purposes.

APPENDIX A: CREATING AN XBRL DOCUMENT

Here, we briefly describe how to create an XBRL document and an XSLT style sheet. Please note that this example is based on version 1.0 of XBRL. The XBRL specification version 2.0 was publicly announced on Dec. 14, 2001. However, at the time this article was prepared, the XBRL taxonomy version 2.0 was not yet available. Version 2.0 is significantly different from version 1.0, but the key concepts described in this article are still applicable. Suppose Waterloo Inc. wishes to prepare its financial statements and distribute them to creditors, investors, analysts, and regulatory parties over the Internet in XBRL format. Table A1 shows the simple balance sheet of Waterloo Inc.

An XBRL document is a collection of elements and attributes that describe financial information and data. The XBRL document can be created by using a text editor (a software application that enables a user to create and modify text files), such as WordPad, or an XML editor or an XBRL instance creation tool. The XBRL document for the balance sheet of Waterloo Inc. is shown below.

Table A1 Balance Sheet

Balance Sheet *Waterloo Inc.*		
(Dollars in thousands)	**2000**	**1999**
Cash and cash equivalents	18,000	15,000
Accounts receivable, net	12,000	5,000
Total assets	30,000	20,000
Accounts payable	10,000	5,000
Common stock	20,000	15,000
Total liabilities and stockholders' equity	30,000	20,000

```
01: <?xml version="1.0" encoding="utf-8"?>
02: <?xml-stylesheet type="text/xsl" href="Appendix.xsl"?>
03:
04: <group xmlns="http://www.xbrl.org/core/xbrl-2000-07-31"
05:        xmlns:ci="http://www.xbrl.org/us/gaap/ci/2000-07-31"
06:        schemaLocation="http://www.xbrl.org/us/gaap/ci/2000-07-31/us-gaap-ci-2000-07-
             31.xsd"
07:        period="2000-12-31"
08:        scaleFactor="3"
09:        precision="9"
10:        type="statements"
```

```
11:         unit="ISO4217:USD"
12:         decimalPattern="#.#">
13:
14:     <!--SECTION: BalanceSheet -->
15:     <group type="ci:statements.balanceSheet">
16:               <group type="ci:cashCashEquivalentsAndShortTermInvestments.cashAndCash
                   Equivalents">
17:                         <label href="xpointer(..)" xml:lang="en">Cash and cash
                             equivalents</label>
18:                         <item period="2000-12-31">18000</item>
19:                         <item period="1999-12-31">15000</item>
20:               </group>
21:               <group type="ci:currentAssets.receivablesNet">
22:                         <label href="xpointer(..)" xml:lang="en">Accounts receivable,
                             net</label>
23:                         <item period="2000-12-31">12000</item>
24:                         <item period="1999-12-31">5000</item>
25:               </group>
26:               <group type="ci:balanceSheet.assets">
27:                         <label href="xpointer(..)" xml:lang="en">Total assets</label>
28:                         <item period="2000-12-31">30000</item>
29:                         <item period="1999-12-31">20000</item>
30:               </group>
31:               <group type="ci:accountsPayableAndAccruedExpenses.accountsPayable">
32:                         <label href="xpointer(..)" xml:lang="en">Accounts payable</label>
33:                         <item period="2000-12-31">10000</item>
34:                         <item period="1999-12-31">5000</item>
35:               </group>
36:               <group type="ci:stockholdersEquity.commonStock">
37:                         <label href="xpointer(..)" xml:lang="en">Common stock</label>
38:                         <item period="2000-12-31">20000</item>
39:                         <item period="1999-12-31">15000</item>
40:               </group>
41:               <group type="ci:balanceSheet.liabilitiesAndStockholdersEquity">
42:                         <label href="xpointer(..)" xml:lang="en">Total liabilities and
                             stockholders' equity</label>
43:                         <item period="2000-12-31">30000</item>
44:                         <item period="1999-12-31">20000</item>
45:               </group>
46:     </group>
47: </group>
```

The first line is the XML declaration: <?xml version = "1.0" encoding = "utf-8"?>. It is an example of an XML processing instruction. The XML processing instruction starts with "<?" and ends with "?>" and the first word after "<?" is the name of the processing instruction. Usually, an XML document starts with the XML declaration that specifies the version of XML being used. Thus version = "1.0" indicates that the document conforms to XML 1.0. The XML declaration may also have several attributes. In the above example, the "encoding" attribute specifies which character encoding is being used. UTF-8 is used for the example document. UTF stands for UCS (Universal Character Set) transformation formats. It is a compressed form of unicode that leaves pure ASCII text unchanged. Therefore, XBRL documents that contain nothing but the common ASCII characters can be edited with applications that do not deal with multibyte character sets like Unicode.

Line 2 contains the processing instruction required to display the XBRL document in the Web browser. The processing instruction, <?xml-stylesheet?>, has two attributes, "type" and "href." The type attribute specifies the style sheet language used, and the href attribute specifies a URL where the style sheet is located. In our example, the type is "text/xsl" and the href is "Appendix.xsl." These are further explained and illustrated later.

XBRL documents are created by using one or more taxonomies. A taxonomy is a dictionary of the financial terms used in preparing business reports, such as financial statements. The taxonomy can be a commonly accepted taxonomy, such as the CI (Commercial and Industrial) taxonomy created by XBRL.ORG, or other taxonomies created by users for their specific requirements. Lines 4 through 12 indicate the XML namespace for the taxonomy used to create the example document, namely the CI taxonomy (us-gaap-ci-2000-07-31.xsd). XML namespaces provide a

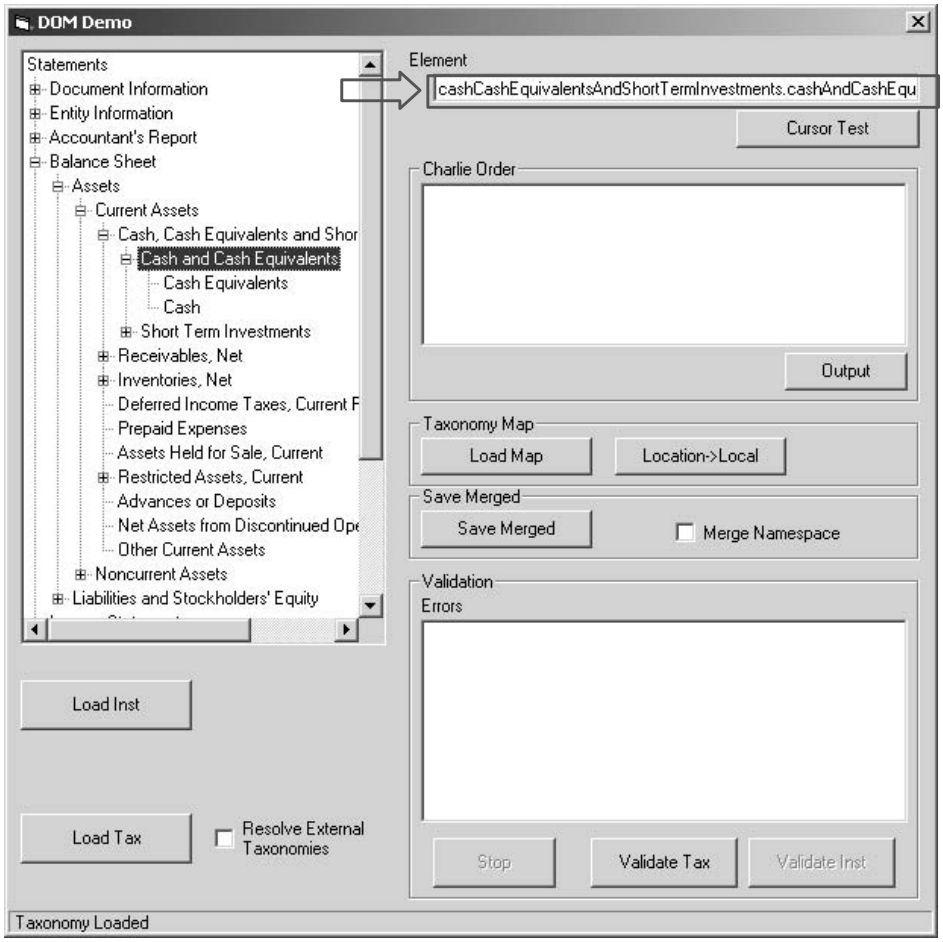


Figure A1: XBRL "Cash and Cash Equivalents" element in a taxonomy viewer application.

simple method for qualifying element and attribute names used in XML documents by associating them with namespaces identified by URL references. Thus, they uniquely identify a set of names so that there is no ambiguity when objects having different origins but the same names are mixed together. In an XML document, any element type or attribute name can thus have a two-part name, consisting of the name of its namespace and then its local (functional) name. (Further information may be found at the W3C Web site, http://www.w3.org.)

The main task in creating an XBRL document is mapping the company's financial information to XBRL financial terms contained in the taxonomy. This involves finding the XBRL elements that correspond to the company's financial information. For example, lines 15 and 46 show the XBRL elements for the balance sheet. The <group type = "ci:statements.balanceSheet"> is a start tag, and </group> is an end tag. The XBRL elements between the two tags describe each balance sheet item.

Because XBRL taxonomy files are XML files, users can find XBRL elements that correspond to their financial information by using a text editor, as illustrated in Figure 9. Although it is possible for users to view taxonomy information with any text editor, a taxonomy viewer application can help users obtain a better and faster understanding of the structure of this information. Therefore, with a taxonomy viewer application, users can easily prepare XBRL documents by mapping XBRL elements in various taxonomies to their financial information. For instance, lines 16 through 20 illustrate the XBRL elements for the cash and cash equivalents item. Using the taxonomy viewer application, users obtain the XBRL element name for their cash and cash equivalents item. A screen shot of the XBRL "Cash and Cash Equivalents" element in a taxonomy viewer application is shown in Figure A1. The taxonomy viewer application used in Appendix I is available from XBRL Solutions Inc., at http://www.xbrlsolutions.com.

The name of the XBRL taxonomy element for cash and cash equivalents is "cashCashEquivalentsAndShortTermInvestments.cashAndCashEquivalents." Thus, the start tag for the cash and cash equivalents item is <group type = "ci:cashCashEquivalentsAndShortTermInvestments.cashAndCashEquivalents">, and the end tag is </group>.

The XBRL elements between the start and end tags describe "label" and "item" information. Line 17 shows the label of the cash and cash equivalents item, and lines 18 and 19 illustrate the amount of cash and cash equivalents for the years 2000 and 1999, respectively.

By following the same steps, the rest of the balance sheet items in the balance sheet of Waterloo Inc. can be mapped to XBRL elements. Lines 21 through 25 show the

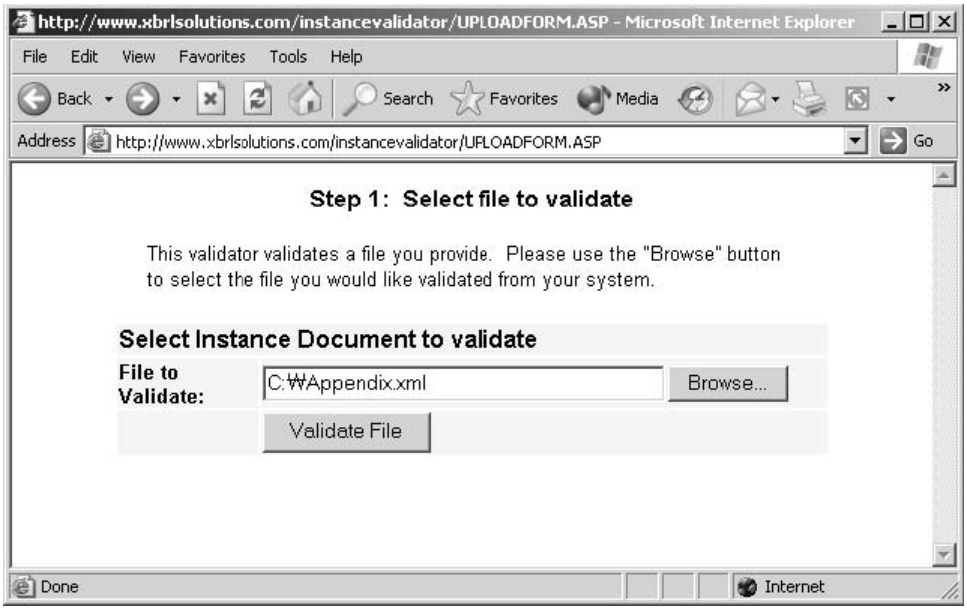


Figure A2: Screen shot of the Internet-based XBRL validation service provided by XBRL Solutions Inc.

XBRL elements for the accounts receivable item, and the XBRL elements for the accounts payable item are shown in lines 31 through 35. Finally, lines 36 though 40 contain the XBRL elements for the common stock item.

Once the XBRL document is created, it is important to check whether the created document is a valid XBRL document. The validation can be performed by using validation software or an Internet-based service. A screen shot of the Internet-based XBRL validation service provided by XBRL Solutions Inc. (http://www.xbrlsolutions.com/) is shown in Figure A2.

An XBRL document provides contextual information, but it does not define how the information should be displayed. To display an XBRL document in a Web browser, users need to write a style sheet (prepared with a style sheet language) for the XBRL document to specify how the document should be displayed. A style sheet can be shared between different documents and different users, as well as integrated with other style sheets. With the appropriate style sheets, users can transform the XBRL document into an HTML document, text document, word processing document, spreadsheet, database file, or another XBRL document. The two most popular style sheet languages are CSS and XSLT. XSLT is currently the only style sheet language designed specifically for use with XML. Thus, for our XBRL example, we will create an XSLT style sheet to transform XBRL into HTML. The corresponding XSLT style sheet example is shown below.

```
01: <?xml version="1.0"?>
02: <xsl:stylesheet xmlns:xsl="http://www.w3.org/1999/XSL/Transform" version="1.0">
03:
04:  <xsl:output method="html"/>
05:  <xsl:template match="/">
06:   <HTML>
07:     <TITLE>XBRL Example</TITLE>
08:     <BODY>
09:      <P ALIGN="CENTER"><B><FONT SIZE="6">Balance Sheet</FONT></B></P>
10:      <P ALIGN="CENTER"><I><B><FONT SIZE="3">Waterloo Inc.</FONT></B></I></P>
11:     <TABLE BORDER="0" CELLPADDING="5" CELLSPACING="0" WIDTH="100%">
12:      <TR>
13:        <TD WIDTH="50%" STYLE="border-bottom-style: solid; border-bottom-width: 3"
            BORDERCOLOR="#000000">
14:          <P ALIGN="LEFT"><FONT SIZE="3">(Dollars in thousands)</FONT></P>
15:        </TD>
16:        <TD WIDTH="25%" STYLE="border-bottom-style: solid; border-bottom-width: 3"
            BORDERCOLOR="#000000">
17:          <P ALIGN="RIGHT"><FONT SIZE="3">2000</FONT></P>
18:        </TD>
```

```
19:        <TD WIDTH="25%" STYLE="border-bottom-style: solid; border-bottom-width: 3"
            BORDERCOLOR="#000000">
20:          <P ALIGN="RIGHT"><FONT SIZE="3">1999</FONT></P>
21:        </TD>
22:      </TR>
23:      <xsl:for-each select="/group/group/group">
24:        <TR>
25:          <xsl:choose>
26:            <xsl:when test="@type='ci:balanceSheet.assets' or @type='ci:balanceSheet.
                 liabilitiesAndStockholdersEquity'">
27:              <TD WIDTH="50%" STYLE="border-top: 2px solid #000000; border-bottom:
                   5px double #000000">
28:                <B><FONT SIZE="3"><xsl:value-of select="label"/></FONT></B>
29:              </TD>
30:              <xsl:for-each select="item">
31:                 <TD WIDTH="25%" STYLE="border-top: 2px solid #000000; border-bottom:
                      5px double #000000">
32:                  <FONT SIZE="3"><P ALIGN="RIGHT">
33:                  <xsl:value-of select="format-number(current(),'###,###')"/>
34:                 </P></FONT></TD>
35:              </xsl:for-each>
36:            </xsl:when>
37:            <xsl:otherwise>
38:              <TD WIDTH="50%" ><FONT SIZE="3"><xsl:value-of select="label"/></FONT></TD>
39:              <xsl:for-each select="item">
40:                 <TD WIDTH="25%"><FONT SIZE="3"><P ALIGN="RIGHT">
41:                   <xsl:value-of select="format-number(current(),'###,###')"/>
42:                 </P></FONT></TD>
43:              </xsl:for-each>
44:            </xsl:otherwise>
45:          </xsl:choose>
46:        </TR>
47:      </xsl:for-each>
48:    </TABLE>
49:   </BODY>
50:  </HTML>
51: </xsl:template>
52:</xsl:stylesheet>
```

Because XSLT uses XML to describe templates, rules, and patterns, an XSLT style sheet starts with the XML declaration <?xml version = "1.0"?>. The XSLT document itself is an "xsl:stylesheet" element. Thus, line 2 of the XSLT style sheet example describes the start of the XSLT process with the XSLT processing instruction: <xsl:stylesheet xmlns:xsl = "http://www.w3.org/1999/XSL/Transform" version = "1.0">

Line 52 indicates the end of the XSLT process: </xsl:stylesheet>. The XSLT processing instruction indicates the XSLT name space—the location of the XML elements that comprise XSLT instructions. XSLT instructions are identified by the "xsl:" prefix on the element. Line 4 specifies the overall method used for outputting the result.

As you can see from the example, an XBRL document has a hierarchical structure whereby the root element is connected to its child elements, each of which may connect to zero or more children of its own, and so forth. Figure A3 displays each element of the XBRL example as a tree structure.

XSLT accepts an XBRL document as input and produces another document as output. An XSLT style sheet consists of a list of templates and instructions. A template has a pattern that specifies the elements it applies to an XBRL document. Thus, when the pattern is matched, instructions contained in the template are applied. In other words, when an XSLT style sheet is applied to an XBRL document, XSLT recognizes the root element of the XBRL document and looks through each child element in turn. As each element in the XBRL document is read, XSLT compares it with the pattern of each template in the style sheet. When XSLT finds an element that matches a template's pattern, it creates outputs by applying the instructions contained in the template.

Each template is represented by using a "xsl:template" processing instruction. The template instruction starts with <xsl:template match = ""> and ends with </xsl: template>. Lines 5 and 51 show an example of a template instruction. Line 5 indicates the template of the root element: <xsl:template match = "/">. Each "xsl:template" instruction has a "match" attribute that specifies which

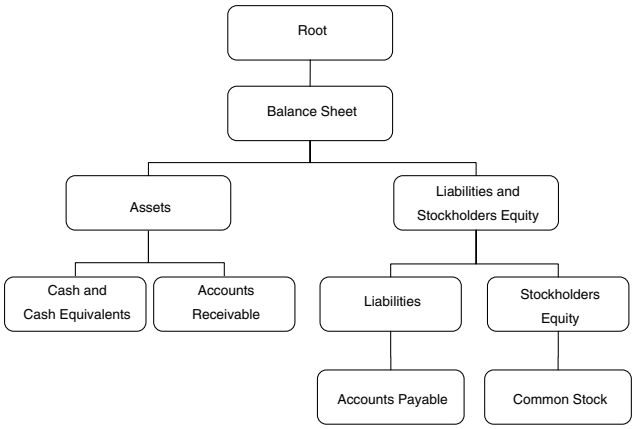


Figure A3: Elements of the XBRL example displayed as a tree structure.

element of the template is instantiated. To specify the root element in the template, the value "/" is given its match attribute.

There are three useful XSLT processing instructions for transforming the value of an element in the XBRL document into an element in the output document. First, the "xsl:value-of" instruction copies the value of an element in the XBRL document into the output document. Each "xsl:value-of" instruction has a "select" attribute that specifies which element's value is being taken. For example, suppose you want to copy the content of the cash and cash equivalents element, specifically the value of its item. You can obtain the value with <xsl:value-of select = "item"/>. The "xsl:value-of" instruction should only be used in contexts where it is unambiguous as to which element's value is being taken, because if there are multiple possible elements that could be selected, only the first one will be chosen.

There are several ways of processing multiple elements in turn. One option is the "xsl:for-each" instruction. The "xsl:for-each" instruction processes each element chosen by its selected attribute in turn. Lines 23 and 47 show an example of the "xsl:for-each" instruction. Both lines illustrate the processing of the child elements of the "group" element. Line 19 shows a start tag <xsl:for-each select = "group/group/group"> and line 54 indicates an end tag </xsl:for-each>. Each "xsl:for-each" instruction has a "select" attribute that specifies which element's value is being taken. To specify the child elements, the value "/" is given its selected attribute. For example, "group/group" indicates that all group elements within a group element should be processed.

Finally, XSLT provides instructions that enable users to process child elements based on the input. One option is the "xsl:choose" instruction. It provides multiple conditional testing in conjunction with the "xsl:when" and "xsl:otherwise" elements. Lines 25 through 45 show an example of the "xsl:choose" instruction. The "test" attribute of "xsl:when" contains a select expression that evaluates to a boolean. Therefore, if the expression is true, the contents of the "xsl:when" instruction are processed. Otherwise, the contents of the "xsl:otherwise" instruction are processed.

Line 25 shows the start tag of the "xsl:choose" instruction: <xsl:choose>. Line 53 shows the end tag: </xsl:choose>. This instruction states that if the "type" attribute of the current element is equal to "ci:statements.balanceSheet" or "ci:balanceSheet.liabilitiesAndStockholdersEquity," then the XSLT instructions between lines 27 and 35 are processed. Otherwise, the XSLT instructions between lines 38 and 43 are processed.

When the XSLT style sheet example is applied to the XBRL document example, the following actions occur.

Lines 1 and 2: These indicate the start of the XSLT process.

Line 4: This XSLT instruction indicates that the results are transformed as HTML.

Line 5: The root element is compared with the pattern of each template in the style sheet. It matches the first one.

Lines 6–22: The following HTML tags are written out.

```
<HTML>
  <TITLE>XBRL Example</TITLE>
  <BODY>
     <P ALIGN="CENTER"><B><FONT SIZE="6"> Balance Sheet</FONT></B></P>
     <P ALIGN="CENTER"><I><B><FONT SIZE="3">Waterloo Inc.</FONT></B></I></P>
     <TABLE BORDER="0" CELLPADDING="5" CELLSPACING="0" WIDTH="100%">
        <TR>
           <TD WIDTH="50%" STYLE="border-bottom-style:solid;border-bottom-width:3"
            BORDERCOLOR="#000000">
              <P ALIGN="LEFT"><FONT SIZE="3">(Dollars in thousands)</FONT></P></TD>
           <TD WIDTH="25%" STYLE="border-bottom-style:solid;border-bottom-width:3"
            BORDERCOLOR="#000000">
              <P ALIGN="RIGHT"><FONT SIZE="3">2000</FONT></P></TD>
           <TD WIDTH="25%" STYLE="border-bottom-style:solid;border-bottom-width:3"
            BORDERCOLOR="#000000">
              <P ALIGN="RIGHT"><FONT SIZE="3">1999</FONT></P></TD>
        </TR>
```

Lines 23–47: The "xsl:for-each" instruction processes each child element of the "group/group/group" element.

Line 24: The HTML tag, <TR>, is written out.

Lines 25–45: The XSLT instruction "xsl:choose" is processed.

Lines 26–36: If the attribute of the current element is equal to "ci:statements.balanceSheet" or "ci:balanceSheet.liabilitiesAndStockholdersEquity," then instructions between lines 27 and 35 are processed. These lines show the XSLT instructions that write out total assets and total liabilities and stockholders' equity.

Line 27: The HTML tag, <TD WIDTH = "50%" STYLE = "border-top: 2px solid #000000; border-bottom: 5px double #000000">, is written out.

Line 28: The value of the current element's label is written out with HTML tags. Example: <B><FONT SIZE = "3"> Total assets </FONT></B>

Line 29: The HTML tag, </TD>, is written out.

Line 30: The "xsl:for-each" instruction processes each "item" element.

Lines 31–32: The <TD WIDTH = "25%" STYLE = "border-top: 2px solid #000000; border-bottom: 5px double #000000"> and <FONT SIZE = "3"><P ALIGN = "RIGHT"> tags are written out.

Line 33: The value of the current element is converted to a formatted number. Example: 30,000.

Line 34: The </P></FONT></TD> tags are written out.

Lines 37–44: Otherwise, instructions between lines 38 and 43 are processed. These lines show the XSLT instructions that write out balance sheet items.

Line 38: The value of the current element's label is written out with HTML tags. Example: <TD WIDTH = "50%"><FONT SIZE = "3"> Cash and cash equivalents </FONT></TD.

Line 39: The "xsl:for-each" instruction processes each "item" element.

Line 40: The <TD WIDTH = "25%"><FONT SIZE = "3"><P ALIGN = "RIGHT"> tags are written out.

Line 41: The value of the current element is converted to a formatted number. Example: 18,000.

Line 42: The </P></FONT></TD> tags are written out.

Line 46: The </TR> tag is written out.

Lines 48–50: The following HTML tags are written out.

```
                </TABLE>
            </BODY>
    </HTML>
```

Lines 51 and 52: These indicate the end of the XSLT process.

After an XBRL document and an XSLT style sheet for that document are created, a processing instruction is required for the XBRL document, to be shown in the Web browser. The processing instruction, <?xml-stylesheet?>, has two attributes, "type" and "href." The type attribute specifies the style sheet language used, and the href attribute specifies a URL where the style sheet is located. Our XBRL example includes an instruction in the XBRL document specifying the following XSLT style sheet to be used: <?xml-stylesheet type = "text/xsl" href = "Appendix.xsl"?>. Currently, the processing instruction <?xml -stylesheet?> is included in the XBRL document to specify the style sheet to be used. However, in the long term, there will be a number of different ways to do it, including browser-server negotiation via HTTP headers,

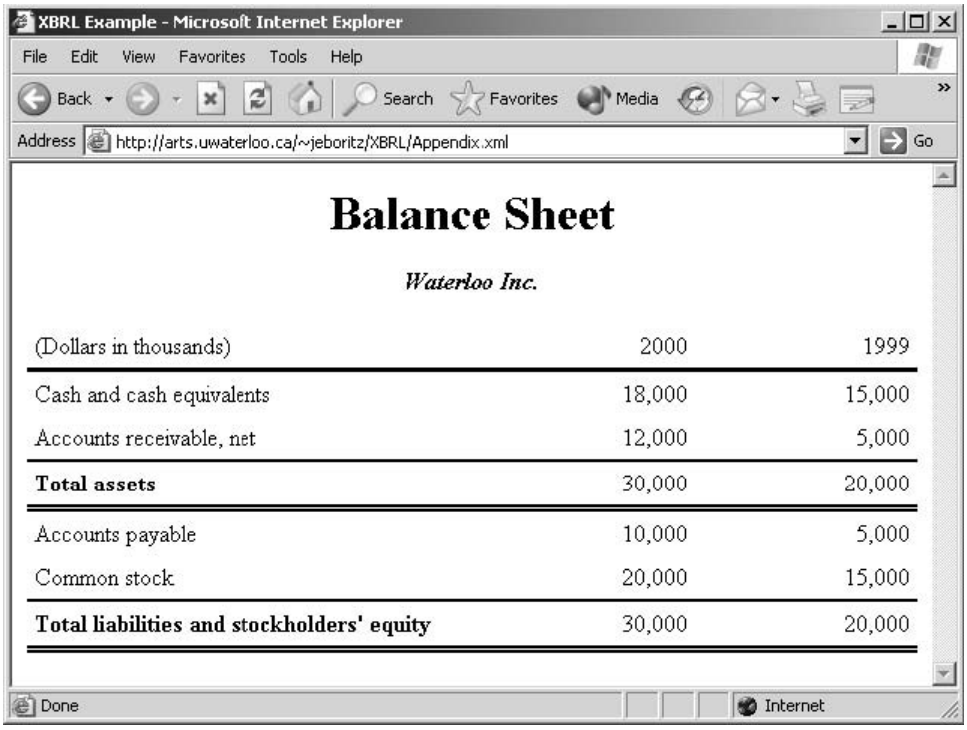


Figure A4: XBRL example.

naming conventions, and browser-side defaults. Figure A4 shows the XBRL example in the Web browser.

Microsoft Internet Explorer 6.0 is recommended for this example. Internet Explorer 5.0 and 5.5 include the Microsoft XML parser (MSXML), which includes an implementation of XSL, which is based on a working draft of the XSLT specification. The XSLT specification was finalized on November 19, 1999. The MSXML parser installed in Internet Explorer 5.0 and 5.5 does not support the most current XSLT specification. Thus, a reader of this article who uses Internet Explorer 5.0 or 5.5 must install MSXML 3.0 or a higher version from http://MSDN.microsoft.com/xml. Unfortunately, Netscape 6.2 does not support the XSLT specification. Therefore, the example will not work in Netscape.

In this appendix, we have explained how to create an XBRL document and an XSLT style sheet. However, in practice, XBRL documents will be created by a software package that automatically creates an XBRL document by mapping a company's financial statements prepared by its internal accounting system to XBRL taxonomy elements and validates the resulting XBRL code. In addition, XSLT style sheets would be created automatically by using a style sheet generating software package. The validated XBRL document can then be made available to users, such as creditors, investors, and analysts, on the company's Web site. Users can analyze the data in the XBRL document by loading it into software, such as spreadsheet and database software, on their own computers.

APPENDIX B: FOR MORE INFORMATION RELATED TO XML AND XBRL

Organization	Web Site
AccountingWEB	http://www.accountingweb.com/
IBM	http://www.ibm.com/xml
KPMG	http://www.kpmg.com
Microsoft	http://msdn.microsoft.com/xml
Microsoft Investor Relations XBRL Information	http://www.microsoft.com/msft/xbrlinfo.htm
NASDAQ	http://www.nasdaq.com/xbrl
OASIS	http://www.oasis-open.org
Oracle	http://technet.oracle.com/tech/xml
PricewaterhouseCoopers	http://www.pwcglobal.com
Sun	http://Java.sun.com/products/xml
The XML Industry Portal	http://www.xml.org
W3C	http://www.w3c.org
W3Schools	http://www.w3schools.com
XBRL.ORG	http://www.xbrl.org
XBRL Educational Resource Center	http://web.bryant.edu/~xbrl/
XBRL Express	http://www.edgar-online.com/xbrl
XBRL Public Discussion	http://www.egroups.com/group/xbrl-public
XBRL Solutions, Inc	http://www.xbrlsolutions.com
XBRL Tools	http://www.xbrl.org/resource_center
XML Files	http://www.xmlfiles.com
XML.COM	http://www.xml.com

GLOSSARY

Attribute A property or characteristic. Color, for example, is an attribute of a person's hair. In using or programming computers, an attribute is a changeable property or characteristic of some component of a program that can be set to different values.

CSS Cascading Style Sheets. It is a means of separating the presentation from the structural markup. Cascading Style Sheets level 1 (CSS1) is a W3C recommendation. It describes the CSS language as well as a basic formatting model. Cascading Style Sheets level 2 (CSS2), which is also a W3C recommendation, builds on CSS1. It includes media-specific style sheets (e.g., printers and aural devices), and element positioning and tables.

DTD Document Type Definition. A DTD defines the tags the document type will use, what they mean, and whether, and to what extent, individual tags can be nested. For example, HTML is a SGML DTD.

Element A concept described by a taxonomy. For example, the element with the name "nonCurrentAssets. propertyPlantAndEquipmentNet" is a concept.

HTML HyperText Markup Language, the basic language for creating a Web page. HTML consists of a set of markup symbols inserted into a file intended for display on a Web browser page. The markup tags tell the Web browser how to display a Web page's words and images for the user.

Instance An XML document containing XBRL elements that together constitute one or more statements. The financial statements of a specific company, expressed in XBRL, would be an instance.

Item A fact reported within a given period of time about a given business entity. Corresponds to an abstract XML element "item" in XBRL.

Markup The sequence of characters or other symbols that are inserted at certain places in a text or word processing file to indicate how the file should look when it is printed or displayed or to describe the document's logical structure. The markup indicators are often called tags.

SGML Standard Generalized Markup Language, an international standard for defining and using document structure and content. SGML incorporates both data labeling and data presentation information but leaves procedural issues entirely to the rendering application.

Style Sheet A mechanism to describe how a document should be displayed. A style sheet is prepared with a style sheet language. Two of the most popular style sheet languages are CSS and XSLT.

Taxonomy A dictionary of the financial terms used in preparing financial statements or other business reports and the corresponding XBRL tags.

XBRL Extensible Business Reporting Language. XBRL is the financial profession's adaptation of XML for financial reporting. A joint industry and government consortium, including the American Institute of Certified Public Accountants (AICPA), six information technology companies, and the five largest accounting and professional services firms, was established for this purpose in the fall of 1999.

XML Extensible Markup Language. XML was invented by adopting the key functions of SGML while excluding the less essential ones. It is extensible because the language can be extended by anyone who wants to create additional tags for new and unforeseen purposes. It is a markup language because XML is a method of tagging information using accepted rules and formats to give definition to text and symbols.

XML Schema An XML-based alternative to a DTD to describe the structure, content and semantics of an XML document.

XSLT Extensible Stylesheet Language Transformations, a style sheet language designed specifically for use with XML. XSLT can transform XML into other documents, such as HTML or database, filter and sort XML data, format XML data, add or remove elements into/from the output file, and rearrange and sort the element.

CROSS REFERENCES

See *Cascading Style Sheets (CSS); Extensible Markup Language (XML); Extensible Stylesheet Language (XSL); HTML/XHTML (HyperText Markup Language/Extensible HyperText Markup Language); Public Accounting Firms*.

REFERENCES

APRA (n.d.). Retrieved April 29, 2003, from http://www.apra.gov.au/

Boritz, J. E., & No, W. G. (2002). *Assurance reporting with XML: XARL (extensible assurance reporting language*). Manuscript, University of Waterloo, Center for Information System Assurance.

Clark, J. (n.d.). Comparison of SGML and XML. Retrieved April 29, 2003, from http://www.w3.org/TR/NOTE-sgml-xml-971215.html

ebXML (n.d.). Retrieved April 29, 2003, from http://www.ebxml.org/

Fin XML (n.d.). Retrieved April 29, 2003, from http://www.finxml.org/

FIX (n.d.). Retrieved April 29, 2003, from http://www.fixprotocol.org/

FpML (n.d.). FpML: The XML standard for swaps, derivatives and structured products. Retrieved April 29, 2003, from http://www.fpml.org/

Microsoft (n.d.). MSFT investor relations. Retrieved April 29, 2003, from http://www.microsoft.com/msft/

Open Financial Exchange (n.d.). Retrieved April 29, 2003, from http://www.ofx.net/

Smith, M. (1996). Complex typography: How an early printer eliminated the scribes' red. *Typography Papers, 1*, 75–92.

Web Design Group (n.d.). Linking style sheets to HTML. Retrieved April 29, 2003, from http://www.htmlhelp.com/reference/css/style-html.html

W3C (n.d.a). XML Linking Language (XLink) version 1.0. Retrieved April 29, 2003, from http://www.w3.org/TR/xlink/

W3C (n.d.b). XML schema. Retrieved April 29, 2003, from http://www.w3.org/XML/schema.html

W3C (n.d.c). Cascading Style Sheets. Retrieved April 29, 2003, from http://www.w3.org/style/css

W3C (n.d.d). Extensible Stylesheet Language (XSL) version 1.0. Retrieved April 29, 2003, from http://www.w3.org/TR/xsl/

XBRL Home Page (2000). Retrieved August, 2001, from http://www.xbrl.org/

XMLEDI (n.d.). Retrieved April 29, 2003, from http://www.xmledi-group.org/

FURTHER READING

Bosak, J. (1997). *XML, Java, and the future of the Web*. Retrieved May 12, 2002, from http://www.xml.com/pub/a/w3j/s3.bosak.html

Bosak, J., & Bray, T. (1999). *XML and the second generation Web*. Retrieved January, 2002, from http://www.sciam.com/article.cfm?articleID=0008C786-91DB-1CD6-B4A8809EC588EEDF

Floyd, M. (1998). *A Conversation with Charles F. Goldfarb*. Retrieved January, 2002, from http://www.webtechniques.com/archives/1998/11/beyo/

Halfhill, T. R. (1999). *XML: the next big thing*. Retrieved February, 2002, from http:// http://domino.research.ibm.com/comm/wwwrthinkresearch.nsf/pages/xml199.html

Harold, E. R. (2001). *XML Bible 2nd Edition*. New York: John Wiley & Sons.

Hoffman, C., Kurt, C., & Koreto, R. J. (1999). *The XML files*. Retrieved January, 2002, from http://www.aicpa.org/pubs/jofa/may1999/hoffman.htm

Hoffman, C., & Strand, C. (2001). *XBRL essentials*. New York: AICPA.

MSDN Library (2001). *XML tutorial*. Retrieved March, 2002, from http://msdn.microsoft.com/library/default.asp?url=/library/en-us/xmlsdk30/htm/xmtutxmltutorial.asp

MSDN online Web Workshop (2001). *XML (extensible markup language)*. Retrieved April, 2002, from http://msdn.microsoft.com/xml/general/index.htm

The CoverPages (1999). *AICPA, Information Technology Companies, and Five Largest Accounting and Professional Service Firms Join Forces in Developing XML-Based Financial Reporting Language*. Retrieved February 18, 2002, from http://xml.coverpages.org/xfrmlAnn.html

Schatz, B. R. (1997). Information retrieval in digital libraries: Bringing search to the net. *Science, 275, 327–334* Retrieved March 11, 2003, from http://www.canis.uiuc.edu/archive/papers/science-irdl-journal.pdf

Schmidt, W. C., & Cohen, E. E. (1999). *A better language for utilizing the Web*. Retrieved February, 2002, from http://www.nysscpa.org/cpajournal/f201199m.html

Watson, L. A., McGuire, B. L., & Cohen, E. E. (2000). *Looking at business reports through XBRL-tinted glasses*. Retrieved April, 2002, from http://www.strategicfinancemag.com/2000/09g.htm

W3C (1999). *XML in 10 points*. Retrieved February, 2002, from http://www.w3.org/XML/1999/XML-in-10-points.html

XBRL Home Page (2000). *Overview/facts sheet*. Retrieved August, 2001, from http://www.xbrl.org/Faq.htm

XBRL Home Page (2000). *Financial reporting for commercial and industrial gompanies, US GAAP*. Retrieved

August, 2001, from http://www.xbrl.org/us/gaap/ci/2000-07-31/us-gaap-ci-2000-07-31.htm

XBRL Home Page (2000). *Extensible business reporting language (XBRL) specification*. Retrieved August, 2001, from http://www.xbrl.org/TR/2000-07-31/XBRL-2000-07-31.htm

XBRL Resource Center (2001). *XBRL educational resource section*. Retrieved November, 2001, from http://web.bryant.edu/~xbrl/index.html

Zarowin, S., & Harding, W. E. (2000). Finally, business talks the same language. *Journal of Accountancy,* August, 24–30.

Reviewers

June M. Abbas State University of New York at Buffalo
Russ Abbott California State University, Los Angeles
Sibel Adali Rensselaer Polytechnic Institute
Carlisle M. Adams Senior Cryptographer, Entrust, Inc.
E. Scott Adler University of Colorado, Boulder
Fred J. Aebli Penn State University
Van B. Afes New York University
Ilieva Ilizastigui Ageenko Wachovia Corporation
Jagan Agrawal Sonoma State University
Manish Agrawal University of South Florida
Gurvinder Ahluwalia Intervector Technologies, Inc.
Gail-Joon Ahn University of North Carolina at Charlotte
Shakil Akhtar Lucent Technologies, Inc.
Michael A. Albert Harvard Extension School
Jonathan Aldrich University of Washington
Roger T. Alexander Colorado State University
Jim Allan Texas School for the Blind and Visually Impaired
Kathleen Allen University of Southern California
Marcus T. Allen Florida Atlantic University
Sheri Alpert University of Notre Dame
Patrick Alphonso Swamiware LLC
Maria Dolores Alvarez Boğaziçi University, Turkey
Gary C. Anders Arizona State University–West
Dorine C. Andrews University of Baltimore
Fedor Andrianov The University of California, Los Angeles
Marios C. Angelides Brunel University, United Kingdom
Edward D. Arnheiter Rensselaer Polytechnic Institute
Michael A. Arnzen Seton Hill University
Alfred W. Arsenault Diversinet Corp.
Angelo Artale University of California, Berkeley
Robert C. Ash University Southeast
Okechukwu B. Asobie University of Colorado at Denver
Giuseppe Ateniese The Johns Hopkins University
Douglas M. Auclair Cotillion Group, Inc.
Elias M. Awad University of Virginia
Mike Axelrod Rochester Institute of Technology
John W. Bagby The Pennsylvania State University
Mark Baker University of Portsmouth, United Kingdom
Paul M. A. Baker Georgia Institute of Technology
Dan Balan PMCV Technologies
Dirk Baldwin University of Wisconsin, Parkside
Narasimhaswamy Banavara Mercy College
Joan Bantz The Evergreen State College
Indranil R. Bardhan The University of Texas at Dallas
Robert N. Barger University of Notre Dame
Joey Bargsten University of Oregon
Lewis Barnett University of Richmond
Anat BarNir University of North Texas
Connie L. Bauer Marquette University
Paul M. Bauer University of Denver
Coskun Bayrak University of Arkansas at Little Rock
Annelise J. Bazar West Valley College
Michael D. Beavers Lake Land College
Greg Becker University of Maryland
Andrew Beckerman-Rodau Suffolk University
Barton Beebe Yeshiva University
Salvatore Belardo State University of New York, at Albany
Colleen Bell University of Oregon
Mary Ann Bell Sam Houston State University
Thomas W. Bennet Mississippi College
Robert Berezdivin George Mason University
Anthony K. Betrus State University of New York at Potsdam
B. Bhagyavati Columbus State University
Harvey Bingham W3C Web Accessibility Initiative
Chatschik Bisdikian IBM T. J. Watson Research Center
Robyn Blakeman West Virginia University
Jody Blanke Mercer University
Julie Bobay Indiana University Bloomington Libraries
Paul Bohman Utah State University
Shawn A. Bohner Virginia Tech
John E. Boon RAND Corporation
Aaron Bor California State University, Chico
Alex Bordetsky Naval Postgraduate School, Monterey, California
Polly D. Boruff-Jones Indiana University Purdue University Indianapolis
William H. Bowers Penn State Berks-Lehigh Valley College
Paul Bracke Arizona Health Sciences Library
Thomas F. Brady Purdue University, North Central
Yale M. Braunstein University of California, Berkeley
Rowland A. Brengle Anne Arundel Community College
Susan W. Brenner University of Dayton
Hank Bromley The State University of New York at Buffalo
Stephen T. Brower Raritan Valley Community College
Barry N. Brown University of Montana
Jeremy M. Brown North Dakota State University
Judith C. Brown University of Memphis
Timothy X. Brown University of Colorado, Boulder
Ray Bruttomesso Bowditch & Dewey, LLP
Maja Bucar University of Ljubljana, Slovenia
Mark Bullen University of British Columbia, Canada
Ulla Bunz University of Kansas
Nicola Burgess University of East Anglia, United Kingdom
Gary Burnett Florida State University
Ray Cafolla Florida Atlantic University
Xiaomei Cai University of Delaware
Metin Çakanyildirim University of Texas at Dallas
Dean S. Caldwell University of Maryland, University College
Dale W. Callahan University of Alabama at Birmingham
Mike Calvo Citronella Software
Rafael A. Calvo The University of Sydney, Australia
Hasan Cam Arizona State University
James F. Campbell University of Missouri–St. Louis
Heidi Campbell University of Edinburgh, United Kingdom
Kim Sydow Campbell University of Alabama
Radu Campeanu York University, Canada
Randy L. Canis University of Missouri–Rolla
James Cannady Nova Southeastern University

Ricardo Capretta University of California at Davis
Andrew Careaga Independent Consultant
Saul Carliner City University of Hong Kong, Hong Kong
David E. Carlson University of Florida
Donald E. Carlson Motorola Corporation
Neil Carlson Duke University
Sandra M. Carriker North Shore Community College
Andy Carvin Benton Foundation
Terri Castaneda California State University, Sacramento
Maria Raquel Garcia Castillo Franklin Pierce College
Daniel C. Castle Hewlett-Packard Company
Cynthia L. Cates Towson University
Kristin Chaffin University of North Carolina at Chapel Hill
Valrie Chambers Texas A & M University–Corpus Christi
Henry Chan The Hong Kong Polytechnic University, Hong Kong
Susy S. Chan DePaul University
Rajarathnam Chandramouli Stevens Institute of Technology
Richard Chapman Auburn University
Kaushal Chari University of South Florida
Bo Chen Hyperion Solutions Corp.
Jim Q. Chen St. Cloud State University
Joyce Chen University of Northern Iowa
Leida Chen Creighton University
Bob R. Cherry University of Central Florida
Ronald L. Chichester South Texas College of Law
Mark J. Chopping Montclair State University
Ken Christensen University of South Florida
Leann Christianson California State University, Hayward
John Chuang University of California at Berkeley
Ping-Tsai Chung Long Island University
Paul Chwelos University of British Columbia, Canada
Cynthia Della Cicalese Marymount University
Hye Clark Albuquerque TVI Community College
Melinda K. Cline University of North Texas
Larry Scott Cline Jr. Southwestern Community College
Eric E. Cohen Price Water House Coopers LLP
Herbert Cohen Webster University
John Coliton Montgomery College
Catherine Collier University of Rochester
J. Stephanie Collins Southern NH University
Sue Conger University of Dallas
Thomas J. Connolly Villanova University
David E Cook University of Derby, United Kingdom
Rebecca Corliss Valdosta State University
Brian Corrie New Media Innovation Centre
Richard R. (Dick) Cottrell University of St. Thomas
J. Philip Craiger University of Nebraska at Omaha
Jeremy W. Crampton Georgia State University
Lorrie Faith Cranor AT&T Labs-Research
Jon Crowcroft University of Cambridge, United Kingdom
Paul E. Cule Marquette University
Allen H. Cunningham Sonoma State University

Christos J. Dabekis Boston College
David Dailey Slippery Rock University
Michael P. D'Alessandro University of Iowa
David T. Damery University of Massachusetts, Amherst
John P. D'Arcy Temple University
R. Dash The University of Texas at Arlington
Luiz A. DaSilva Virginia Polytechnic Institute and State University
Jaime Davila Hampshire College
Champ Davis Chicago-Kent College of Law
Diana S. Dawson Florida Atlantic University
Michael J. Day Northern Illinois University
John Debenham University of Technology, Sydney
Sabrina De Capitani di Vimercati Università di Brescia, Italy
Michele Decker Sypris Electronics, LLC
Linda deLeon University of Colorado at Denver
Lynn A. DeNoia Rensselaer (Polytechnic Institute) at Hartford (CT)
Francis De Vericourt Duke University
Subhankar Dhar San Jose State University
Nik Dholakia University of Rhode Island
Thomas W. Dillon James Madison University
Wei Ding University of Houston–Clear Lake
Lisa Cingiser DiPippo University of Rhode Island
Eck Doerry Northern Arizona University
Michael Doherty University of the Pacific
Hans-Peter Dommel Santa Clara University
Jin Song Dong National University of Singapore, Singapore
Karen Donohue University of Minosota
Rich Dorfman Waukesha County Technical College
Constantinos Dovrolis University of Delaware
Judy Druse Washburn University
Vitaly Dubrovsky Clarkson University
Tom Duncan University of Colorado, Boulder
John L. DuPré Hamilton, Brook, Smith & Reynolds, P.C.
William H. Dutton University of Oxford, United Kingdom

Rick Eary Communications Advisory Group, Inc.
Thomas A. Easton Thomas College
John L. Eatman University of North Carolina at Greensboro
Terrence R. Edwards Central College
George Efthymoglou University of Pireas, Greece
Birgul Egeli Bogazici University, Turkey
Roger Ehrich Virginia Tech
Thomas A. Ehrich Golden Gate University
Wafa Elgarah University of Central Florida
Larry Elin Syracuse University
Osama Eljabiri New Jersey Institute of Technology
Heidi J. C. Ellis Rensselaer (RPI) at Hartford
David Ely San Diego State University
Magda El Zarki University of California, Irvine
Bob Emiliani Rensselaer Polytechnic Institute
Nathan L. Ensmenger University of Pennsylvania
Linda H. Espey Drake University
David Evans University of Virginia
Ray Everett-Church ePrivacyGroup.com
Ezegozie Eze DezeC (Worldwide)

Richard E. Fairley Oregon Graduate Institute
Xiao Fang The University of Arizona
Christine Haight Farley American University
Sarah P. Farrell University of Virginia
Kurt D. Fenstermacher University of Arizona
Robert G. Fichman Boston College
Cliff Figallo Independent Consultant
Raphael Finkel University of Kentucky
Martin Fischer Stanford University
Susanna Frederick Fischer The Catholic University of America
Mary Ann Fitzgerald University of Georgia
Marcia H. Flicker Fordham University

Adina Magda Florea Worcester Polytechnic Institute
Renée Florsheim Loyola Marymount University
Joel Foreman George Mason University
Andres Fortino George Mason University
Patrick N. Foster Central Connecticut State University
Tom Foster Boise State University
William Foster Arizona State University–West
Elizabeth Fraas Eastern Kentucky University
Dennis J. Frailey Raytheon Company
Matt Franklin University of California at Davis
John Freebairn The University of Melbourne, Australia
Allan Fromen Market Intelligence for Sales & Distribution
Anthony Y. H. Fung The Chinese University of Hong Kong, Hong Kong

Patrick Shane Gallagher George Mason University
Dennis F. Galletta University of Pittsburgh
John E. Galvin Indiana University, Indianapolis
Jagdish S. Gangolly State University of New York at Albany
Denise A. Garofalo SUNY Albany SISP
Kelly Garrett University of Michigan
Laura N. Gasaway DePaul University
Susan Gauch University of Kansas
William C. Giauque Brigham Young University
Llewellyn Joseph Gibbons University of Toledo
Brian J. Gibson Auburn University
John Gilbert Digital Visions, Inc.
Mary C. Gilly University of California, Irvine
Maria Gini University of Minnesota
Roy J. Girasa Pace University
Robert L. Glueckauf University of Florida
Eric Goldman Marquette University
Thomas J. Goldsby The Ohio State University
Yun Gong Widener University
Suresh Gopalakrishnan Rutgers University
Steven R. Gordon Babson College
Saurabh Goyal University of California at Davis
Allan W. Graham University of Rhode Island
Andrew Michael Gravell University of Southampton, United Kingdom
Marilyn Greenstein Arizona State University–West
Dawn G. Gregg University of Colorado at Denver
John L. Gresham Fontbonne University
Ronald E. Grieson University of California, Santa Cruz
Louise Gross Cameron University
Christine Guenette St. Louis Community College at Meramec
Stacy J. Guffey Southwestern Community College
Mohsen Guizani Western Michigan University
Jeff H. Gullion Des Moines Area Community College
T. Aaron Gulliver University of Victoria, Canada
Fikret Gürgen Boğaziçi University, Turkey

Sean M. Hackett Vanderbilt University
Creighton Hager Virginia Polytechnic Institute and State University
Thomas Haigh The University of Pennsylvania
Alexander Halavais State University of New York at Buffalo
Debra Haley American University of Paris
Jennifer Hall Massachusetts College of Art
Diane Hamilton Rowan University
Robert J. Hammell Towson University
Joe B. Hanna Auburn University
Christopher P. Harding Canyon College
Ronald Kane Hardy Central College–Geisler Library
Douglas Harris Marquette University
Ernan Haruvy University of Texas at Dallas
Paul G. Harwood University of Maryland
Waël Ali Hassan University of Ottawa, Canada
David Hawking CSIRO Mathematical and Information Sciences, Australia
Richard P. Hayes McGill University, Canada
Helen Hayes Wykle University of North Carolina at Asheville
Elizabeth Haynes University of Southern Mississippi
Sunil Hazari University of Maryland, College Park
Lenwood S. Heath Virginia Tech
Robert W. Heath Jr. The University of Texas at Austin
Jeff Heflin Lehigh University
Vish Hegde University of San Francisco
P. Bryan Heidorn University of Illinois at Urbana–Champaign
Rodney J. Heisterberg Notre Dame de Namur University
Markus Helfert University St. Gallen, Switzerland
Rachelle S. Heller The George Washington University
Linda Hemenway Santa Rosa Junior College
Terrence Hendershott University of California at Berkeley
Joel Henry The University of Montana
Katherine G. Herbert New Jersey Institute of Technology
David B. Heroy The University of Texas at Dallas
Debbie Herring University of Sheffield
Eileen Herring University of Hawaii at Manoa
George A. Heyman Oakton Community College
Kenneth Einar Himma University of Washington
Eric von Hippel Massachusetts Institute of Technology
Hossam H. H'mimy Ericsson Inc.
Robert Hofkin Goldey-Beacom College
James R. Holt Washington State University
Terri L. Holtze University of Louisville
Matthias Holweg Massachusetts Institute of Technology
Brian T. Howard Bridgewater College
Junling Hu University of Rochester
Allison Hunter Web Design, LLC

Seung Bae Im California State University, Chico
Sam Inala University of California, San Diego
Laura R. Ingraham North Carolina State University
James A. Inman University of South Florida
Laurent Itti University of Southern California
Jakob Holden Iversen University of Wisconsin, Oshkosh
Lakshmi Iyer The University of North Carolina at Greensboro

William K. Jackson Southern Oregon University
Andrew T. Jacobs SUNY Rockland Community College
Anne R. Jacobs The Emory University
Eliot Jacobson University of California, Santa Barbara
Charles W. Jaeger Southern Oregon University
Sushil Jajodia George Mason University
Markus Jakobsson RSA Laboratories
Anil B. Jambekar Michigan Technological University
Jeremiah W. James University of Kansas
Krishna Jayakar The Pennsylvania State University
Mark Jeffery Northwestern University, Kellogg School of Management
James Paul Jenal Loyola Law School
Deanna Jenness First Call Computer Solutions
Carol M. Jessup University of Illinois at Springfield
Jesper M. Johansson Microsoft Corporation
Terri Lynn Johnson Indiana University
Matthew Jones University of Cambridge, United Kingdom

Nory B. Jones University of Maine
W. T. Jordan Texas A&M University–Texarkana
Paul Juell North Dakota State University
Ari Juels RSA Laboratories

Joseph Kabara University of Pittsburgh
Yasmin B. Kafai University of California at Los Angeles
James Kalmbach Illinois State University
Manjunath Kamath Oklahoma State University
Philip Kaminsky University of California at Berkeley
Kevin Kane Iowa State University
Peter L. Kantor Hudson Valley Community College
Jahangir Karimi University of Colorado at Denver
Erin Karper Purdue University
Gerald E. Karush Southern New Hampshire University
James Katzenstein California State University, Dominguez Hills
Andrea Lee Kavanaugh Virginia Tech
Charles D. Kay Wofford College
Joseph M. Kayany Western Michigan University
Perry Keller King's College London
George Kelley Morehead State University
Stephen Kent BBN Technologies
Gary C. Kessler Champlain College
Mohammed Ketel North Carolina A&T State University
Sandra J. Kiehl Linfield College
John M. Kiener William Rainey Harper College
Bonn-Oh Kim Seattle University
Dan Jong Kim Michigan State University
Soo Dong Kim Soongsil University, Korea
Sara E. Kimball University of Texas, Austin
Irwin King The Chinese University of Hong Kong
Robin King Georgetown University
Russell E. King North Carolina State University
Joseph M. Kizza University of Tennessee, Chattanooga
Howard J. Klein The Ohio State University
Brad Kleindl Missouri Southern State College
Kurt H. Knudsen Lombard, Knudsen & Holtey
Charles D. Knutson Brigham Young University
Eun-Joung Ko Marymount University
Ann Hibner Koblitz Arizona State University
Thomas R. Kochtanek University of Missouri
Wallace Koehler Valdosta State University
Mark A. Kon Boston University
Kevin Kornegay IBM T. J. Watson Research Center
Michael F. Kosloski Old Dominion University
Jim Krause Indiana University, Bloomington
Malini Krishnamurthi California State University, Fullerton
Prashant Krishnamurthy University of Pittsburgh
Philip S. Kronenberg Virginia Polytechnic Institute and State University
Nir Kshetri University of Rhode Island
Harumi Kuno Hewlett-Packard Laboratories
Ojoung Kwon California State University, Fresno

Deborah LaBelle Penn State Delaware County
Karim Lakhani Massachusetts Institute of Technology
Roberta Lamb University of Hawaii, Manoa
Cristina Lammers South Dakota State University
Douglas Lamont DePaul University
Michael Landry Northeastern State University
Phillip A. Laplante Penn State University
Gail Ann Lasprogat Seattle University
Daniel L. Lau University of Kentucky
Konstantin Läufer Loyola University Chicago
Christine Lavelle Arizona Interactive Media Group
Rob Law Hong Kong Polytechnic University, Hong Kong
Rita L. Laxton Oregon Health and Science University
Karen S. Layne University of Nevada, Las Vegas
John H. Le Kendall Campus Microcomputer Institute
Sharon Lechter Cashflow Technologies, Inc.
Jack Baldwin LeClair Montclair State University
Jae Kyu Lee Korea Advanced Institute of Science and Technology, Korea
Michelle Lee Verisign, Inc.
Raymond Lee Hong Kong Polytechnic University, Hong Kong
Stewart Andrew Leech The University of Melbourne, Australia
Walter S. Leipold Penn State University
Edward M. Lenert City University of New York
Margarita Maria Lenk Colorado State University
Lawrence M. Lesser University of Maryland
Allen H. Levesque Center for Wireless Information Network Studies, Worcester, MA
Vincent LeVeque Science Applications International Corporation
Norman S. Levine Bowling Green State University
Irvin Jay Levy Gordon College
Edwin E. Lewis, Jr. The Johns Hopkins University
Teresita S. Leyell Washburn University
Yao Liang Virginia Polytechnic Institute and State University
Stephen D. Lichtenstein Bentley College
Yossi Lichtenstein University College Dublin, Ireland
Thom Lieb Towson University
Alvin Lim Auburn University
Kwanghui Lim National University of Singapore, Singapore
Carolyn A. Lin Cleveland State University
T. Y. Lin San Jose State University
William Lin Purdue University at Indianapolis
Xia Lin Drexel University
Marlyn Kemper Littman Nova Southeastern University
Bing Liu University of Illinois at Chicago
Ling Liu Georgia Institute of Technology
Mei-Ling Liu California Polytechnic State University, San Luis Obispo
Peng Liu The Pennsylvania State University
Arthur Lizie Bridgewater State College
Barbara Lloyd Independent Consultant
Lanny Lockhart Rochester Institute of Technology
Ronald Lodewyck California State University, Stanislaus
Bruce M. Logan Lesley University
Joyce P. Logan University of Kentucky
Michael W. Longan Valparaiso University
Bob Louden The City University of New York
Dianne B. Love University of Houston, Clear Lake
Kara Lovett Iowa State University
Robert H. Lowson University of East Anglia, UK
Susan Lucas The University of Alabama
Rhonda R. Lummus Iowa State University

Zakaria Maamar Zayed University, United Arab Emirates
Juliana Maantay City University of New York
Mark Mabrito Purdue University, Calumet Hammond
Robert S. MacArthur Boston University
Bertrum H. MacDonald Dalhousie University, Canada

Laurie E. MacDonald Bryant College
Vijay Machiraju Hewlett Packard Laboratories
Fiona C. Maclachlan Manhattan College
Bruce MacLeod University of Southern Maine
Fred H. Mader Marshall University
Grerory R. Madey University of Notre Dame
Michael Madison University of Pittsburgh
Gregory M. Magnan Seattle University
Mary Lou Maher University of Sydney, Australia
Alison Mainwood King's College London, United Kingdom
William P. Manion The University of Maine
Andrew Manison Ansel Systems, Inc.
Richard D. Manning Nova Southeastern University
Joe Marchal James Madison University
Marlow J. Marchant Eastern Kentucky University
David M. Marcovitz Loyola College in Maryland
Todd Margolis Columbia College Chicago
Julie R. Mariga Purdue University
Zdravko Markov Central Connecticut State University
Normand M. Martel Medical Technology Research Corp.
Prabhaker Mateti Wright State University
Anish Mathuria University of Massachusetts, Dartmouth
Rom Mattesich GlobalXchange Communications, Inc
Florian Matthes Technical University Munich, Germany
Margery Mayer Strategy & Process Experts (SPE)
Kelli McCormack Brown University of South Florida
Patrick McDaniel AT&T Labs
Jeff McDougall Texas A&M University
James V. McGee Kellogg School of Management, Northwestern University
Edward L. McGlone Emporia State University
Wayne V. McIntosh University of Maryland
Alison McKay University of Leeds, United Kingdom
Timothy S. McLaren Washington State University
Mary J. Meixell George Mason University
David Mendonça New Jersey Institute of Technology
Peter S. Menell University of California at Berkeley
Rebecca Mercuri Bryn Mawr College
Walter Merrick Three Rivers Community College
Paul F. Merrill Brigham Young University
Gerald A. Merwin Valdosta State University
Jim Metzler Ashton, Metzler & Associates
Vicki H. Micheletto The University of Montana
Scott F. Midkiff Virginia Tech
Mikhail Mikhailov Worcester Polytechnic Institute
Clark A. Miller University of Wisconsin-Madison
Ethan L. Miller University of California, Santa Cruz
Fred Miller Murray State University
F. Thornton Miller Southwest Missouri State University
Gary E. Miller The Pennsylvania State University
Holmes Miller Muhlenberg College
Norm Miller University of Cincinnati
David L. Mills University of Delaware
Jimmy Mills IBM Corporation
Ali Mir Farooq Virginia Tech
Gerald Mitchell University of Colorado, Boulder
Jennie L. Mitchell Saint Mary-of-the-Woods College
Bamshad Mobasher DePaul University
Kenrick J. Mock University of Alaska, Anchorage
Kenneth L. Modesitt Indiana University–Purdue University
Susan D. Moisey Athabasca University, Canada
Torin Monahan Rensselaer Polytechnic Institute
Frank Montabon Iowa State University
Todd K. Moon Utah State University
William L. Moore University of Utah
Vincent Mor Brown University
Keith A. Morneau Northern Virginia Community College
Susan M. Mudambi Temple University
Waleed A. Muhanna The Ohio State University
Kimberly S. Muma Ferris State University
Albert Muniz DePaul Univeristy
Jamie Murphy The University of Western Australia
Kyle B. Murray University of Alberta, Canada
F. Murtagh Queen's University, Belfast
Mirza B. Murtaza Middle Tennessee State University
B. P. S. Murthi The University of Texas at Dallas
Barry L. Myers Northwest Nazarene University
Brad A. Myers Carnegie Mellon University
Gerald M. Myers Pacific Lutheran University
Robert Myre New Jersey Institute of Technology
Kurt Edward Nalty Austin Community College
William D. Nance San Jose State University
Sridhar Narasimhan Georgia Institute of Technology
Leonard Nass New Jersey City University
Barrie R. Nault University of Calgary, Canada
Carlos J. Navarrete California State Polytechnic University, Pomona
Bruce H. Nearon J.H. Cohn LLP
Dave Netherton Old Dominion University
Gregory B. Newby University of North Carolina at Chapel Hill
Dat-Dao Nguyen California State University, Northridge
Dung "Zung" Nguyen Rice University
Nhut Nguyen The University of Texas at Dallas
John M. Nicholas Loyola University Chicago
Erik Nilsen Lewis & Clark College
Peng Ning North Carolina State University
Joseph F. Norton Norton Solutions Group
Christine Neylon O'Brien Boston College

Brian M. O'Connell Central Connecticut State University
David E. O'Gorman University of Illinois at Springfield
Haemoon Oh Iowa State University
Margaret T. O'Hara East Carolina University
Steve O'Keefe Tulane University College
Michael D. Oliver Bowie & Jensen, LLC
David L. Olson University of Nebraska
Robert F. Otondo The University of Memphis
John E. Ottaviani Edwards & Angell, LLP
Terence T. Ow University of Notre Dame
David A. Owens University of Colorado, at Colorado Springs

Phillip B. Palajac Walsh College of Accountancy and Business Administration
E. Kent Palmer MacMurray College
Prashant Palvia The University of North Carolina at Greensboro
Edward L. Parrish Cabrillo College
Andrea C. Peach Georgetown College
David Pearson Santa Rosa Junior College
Witold Pedrycz University of Alberta, Edmonton, Canada
Adrian Peever Barry University
Jian Pei Simon Fraser University, Canada
Jian Pei State University of New York at Buffalo

Jorge Pérez Kennesaw State University
William Perrizo North Dakota State University
David Petrie California State University, Fullerton
Hassan Peyravi Kent State University
Jack Lazarre Pharm University of Southern California
David J. Phillips University of Texas–Austin
John E. Phillips Mansfield University of Pennsylvania
Thomas Pigg Jackson State Community College
Ronald E. Pitt Bryant College
Terry Plum Mount Holyoke College
Sandra Poindexter Northern Michigan University
Steven E. Poltrock The Boeing Company
W. Benjamin Porr George Mason University
James E. Porter Michigan State University
David G. Post Temple University
Thomas A. Powell University of California, San Diego
Frederick Pratter University of Montana
Andrew Prestage Kern County Superintendent of Schools
Philip R. Prins Seattle Pacific University
N. Todd Pritsky Hill Associates, Inc.
Janice C. Probst University of South Carolina
Darren Prokop University of Alaska, Anchorage
Jyrki Pulkkinen University of Oulu, Finland
J. Mark Pullen George Mason University
John Punin Rensselaer Polytechnic Institute
James M. Purtilo University of Maryland, College Park

Hairong Qi University of Tennessee
Susan M. Quade Canyon Collge of Idaho
Liam Quin W3C XML Activity Lead
Eugene R. Quinn Temple University
Philip Quinn Quinn Interactive, Inc.

Chuck Rachlis University of Toronto, Canada
Mansour Rahimi University of Southern California
Gary Randolph Purdue University, Anderson, Indiana
Arvind Rangaswamy Penn State University
Segeda Ranjeet Troutman Sanders LLP
H. R. Rao The University of New York at Buffalo
Sury Ravindran Arizona State University
C. V. Ravishankar University of California–Riverside
Amy W. Ray Bentley College
Russ Ray University of Louisville
Alan I. Rea Western Michigan University
Lawrence C. Reardon University of New Hampshire
David Reavis Texas A&M University–Texarkana
B. J. Reed University of Nebraska at Omaha
Edward M. Reeve Utah State University
Hue Tran Reynolds Florida Institute of CPAs
Hassan Reza University of North Dakota
Sheryne M. Richardson Clayton College State University
Thomas Rienzo Western Michigan University
Donn Ritchie San Diego State University
Gary O. Roberts Alfred University
G. Keith Roberts University of Redlands
William B. Robinson Harvard University
David Rock The University of Mississippi
Claudia Roda American University of Paris
Joachim Rosenthal University of Notre Dame
Brian J. Rosmaita Hamilton College
Steven C. Ross Western Washington University
Gustavo H. Rossi LIFIA-Informatica, UNLP, Argentina
Kenneth G. Rossi Hawaii Pacific University
Gerald Roth Gonzaga University
Neil C. Rowe U.S. Naval Postgraduate School
Bradley S. Rubin University of St. Thomas
Ingrid Russell University of Hartford
Christopher J. Ruth Six Continents Hotels
Patricia A. Ryan Colorado State University
Martin Ryder University of Colorado at Denver
Kyung D. Ryu Arizona State University

Warren Sack University of California, Santa Cruz
Akhil Sahai Hewlett-Packard Laboratories
Sunil G. Samanta Mercy College
Pierangela Samarati Università di Milano, Italy
Michael Sauers Bibliographical Center for Research (BCR)
Vicki L. Sauter University of Missouri–St. Louis
Akbar M. Sayeed University of Wisconsin
Mike Schiff Current Analysis Inc.
Mark A. Schlesinger University of Massachusetts, Boston
Mark Schmalz University of Florida
Steffen W. Schmidt Iowa State University
Eric Schneider Dakota State University
Ryan Schneider Troutman Sanders LLP
Hans J. (Jochen) Scholl State University of New York at Albany
Hans J. Scholl University of Washington
Jonathan Schroeder Royal Institute of Technology, Stockholn
John H. Schuh Iowa State University
Michael H. Schuldes Dakota State University
Alice Schumaker University of Nebraska at Omaha
Charles M. Schweik University of Massachusetts, Amherst
John A. Scigliano Nova Southeastern University
Leslie A. Scott Microsoft Corporation
Wendy Seltzer St. John's University
Sanjeev Setia George Mason University
F. Stan Settles University of Southern California
Tom Seymour Minot State University
Mark Shacklette University of Chicago
Sylvia Shafto Notre Dame de Namur University
Benjamin B. M. Shao Arizona State University
Susan J Shapiro Indiana University East
Srinarayan Sharma Oakland University
Brett S. Sharp University of Central Oklahoma
Neal G. Shaw Texas Christian University
Shashi Shekhar University of Minnesota
Ron Shelby eFusionSolutions, LLC
Stewart N. T. Shen Old Dominion University
Prashant Shenoy University of Massachusetts, Amherst
Steven Shepard Shepard Communications Group, LLC
Richard C. Sherman Miami University
Zhengzhong Shi North Dakota State University
Sang C. Shin Sun Microsystems
Kristi Siegel Mount Mary College
Tom Siep Bluetooth SIG, Inc.
Judith C. Simon The University of Memphis
Ronald R. Sims College of William and Mary
Vic Sims Southern Oregon University
Elizabeth Sisley University of Minnesota
Matt Slaybaugh New York University
Anne Marie Smith La Salle University
Gregory L. Smith Kansas State University

James M. Smith Hamilton, Brook, Smith & Reynolds, P.C.
Lloyd Smith Southwest Missouri State University
Mark Smith Purdue University, North Central
Stephen M. Smith The Johns Hopkins University
Susan J. Smith Harvard University
William L. Smith Emporia State University
Eli M. Snir Southern Methodist University
Charles Snow Lexiconix Systems
Daniel J. Solove Seton Hall Law School
Changwon Song Drew University
Hongjun Song The University of Memphis
Yvonne Sor Ashenden & Sor
David E. Sorkin The John Marshall Law School
Anna O. Soter The Ohio State University
Victor E. Sower Sam Houston State University
Albert D. Spalding Wayne State University
Stuart M. Speedie University of Minnesota
Dawn E. Spencer University of Southern Colorado
Erich Spencer University of Baltimore
Thomas F. Stafford University of Memphis
Bernd Carsten Stahl University College Dublin, Ireland
Allen Stairs University of Maryland
Linda L. Stanley Our Lady of the Lake University
Kay M. Stanney University of Central Florida
Kenneth R. Stanton University of Baltimore
Stam Whitlock Stathis CPA Associates
Ronald W. Staudt Illinois Institute of Technology
David J. Steele Loyola Law School
Steven J. Steinberg Humboldt State University
Carl Steiner University of Illinois at Chicago
Michael Stiber University of Washington, Bothell
Anna E. Story Mercy College
Troy J. Strader Iowa State University
Vincent Peter Stulginskis Appalachian State University
Shuang Sun The Pennsylvania State University
Xian-He Sun Illinois Institute of Technology
David R. Surma Indiana University, South Bend
Robert J. Sweeney Wright State University
David X. Swenson The College of St. Scholastica
Sharon W. Tabor Boise State University
Hossein Tahani New Mexico Highlands University
Serafin Talisayon University of the Philippines, Philippines
Zixiang Tan Syracuse University
Michael Tang University of Colorado, at Denver
Bob Tarr University of Maryland
Stephen R. Tate University of North Texas
Steven L. Taylor Troy State University
David Tegarden Virginia Tech
Rahul Telang Carnegie Mellon University
Steven F. Tello University of Massachusetts, Lowell
Jack Templin New York University
Annette Marie Tetmeyer University of Kansas
C. S. Thachenkary Georgia State University
James W. Thatcher Accessibility Consulting
Dale R. Thompson University of Arkansas
Gary Thompson Saginaw Valley State University
Stephen W. Thorpe Neumann College
G. D. Thurman Scottsdale Community College
Zhi Tian Michigan Technological University
Dallen J. Timothy Arizona State University
Bruce P. Tis Simmons College
Kerem Tomak University of Texas at Austin
Goran Trajkovski Towson University
Quoc-Nam Tran Lamar University
Issa Traore University of Victoria, Canada
Ming-Hsiang Tsou San Diego State University
Francis D. Tuggle Chapman University
Dan Turk Colorado State University
Catherine Turley George Washington University
Joanne M. Twining Intertwining.org Corporation

Okechukwu C. Ugweje The University of Akron
MacDonnell Ulsch Janus Risk Management, Inc.

Jacques Vaisey Simon Fraser University, Canada
Philip A. Van Allen Art Center College of Design
Hubert B. Van Hoof Northern Arizona University
Pamela Van Hook Devry University–Dallas
Nancy Van House University of California, Berkeley
Jorge Vasconcelos Santillán Johns Hopkins University
Vasja Vehovar University of Ljubljana, Slovenia
Jennifer R. Veltsos Minnesota State University, Mankato
Saipradeep Venkatraman University of Cincinnati
David Vest Colorado State University
Dennis Viehland Massey University
M. S. Vijay Kumar Massachusetts Institute of Technology
Charles L. Viles RateIntegration, Inc.
Michael S. Visser University of Oregon
N. Viswanadham The National University of Singapore, Singapore
William J. Vitucci The George Washington University
Agnès Voisard Freie Universitaet Berlin, Germany
Linda Volonino Canisius College
David Voltmer Rose-Hulman Institute of Technology

Ruth Wallach University of Southern California
Janice R. Walker Georgia Southern University
David C. Wang Verizon Communications
Andy Ju An Wang Southern Polytechnic State University
Dana Ward Pitzer College
David Ward Capitol College
Lynn Ward The Indiana Higher Education Telecommunication System
Sidne Gail Ward University of Missouri–Kansas City
Jack H. Warner University of Oklahoma
Warren C. Weber California State Polytechnic University, Pomona
Thomas C. Webster Long Island University
Jerry C. Wei University of Notre Dame
Charles Weibel University of New Jersey
Bruce D. Weinberg Bentley College
George M. Weinberger Southwest Texas State University
Shelly Weinig Colombia University
Shi Weisong Wayne State University
Anthony Wensley University of Toronto, Canada
Steven Wernikoff John Marshall Law School
Jennifer L. Westerhaus Troutman Sanders LLP
Christopher Westley Jacksonville State University
Yun-Oh Whang University of Central Florida
Jim Whitehead University of California, Santa Cruz
A. E. Whiteing The University of Huddersfield, United Kingdom
Larry Whitman Wichita State University

Dave Whitmore Champlain College
Roger Whitney San Diego State University
George R. Widmeyer University of Michigan
David Charles Wierschem Texas A&M University
Garland D. Wiggs Radford University
Ian F. Wilkinson University of New South Wales, Australia
Denver R. E. Williams University of Central Florida
Jeffrey B. Williams Concordia University Wisconsin
Geoff Willis University of Central Oklahoma
Paul L. Witt University of Texas at Arlington
Peter Wolcott University of Nebraska at Omaha
Mary Wolfinbarger California State University, Long Beach
Dave Wolkowitz MarketSting
Jennifer J. Wood Scripps College
George L. Wooley University of Maryland Eastern Sore
Scott E. Worthington Rochester Institute of Technology
D. J. Wu Drexel University
Liangfu Wu Village of Downers Grove
Finn Wynstra Erasmus University Rotterdam, Netherlands

Yang Xiao The University of Memphis
Li Xiao Michigan State University
Yike Xu University of Iowa
Susan E. Yager Southern Illinois University, Edwardsville
Qi Yang University of Wisconsin–Platteville
Fredrik Ygge Trade Extensions
C. Lwanga Yonke Aera Energy LLC
Mike Young University of Connecticut

Michael J. Zagurek FileTek, Inc.
Leigh E. Zeitz University of Northern Iowa
David Zemmels University of Utah
Weining Zhang University of Texas at San Antonio
Yanqing Zhang Georgia State University
Kaimei Zheng University of Massachusetts, Amherst
Nicholas Zsifkov York University, Canada

Index